Education and Information Technologies (2023) 28:905–971
https://doi.org/10.1007/s10639-022-11152-y

Educational data mining to predict students' academic performance: A survey study

Saba Batool[1] · Junaid Rashid[2] · Muhammad Wasif Nisar[1] · Jungeun Kim[3] · Hyuk-Yoon Kwon[4] · Amir Hussain[5]

Received: 27 November 2021 / Accepted: 2 June 2022 / Published online: 9 July 2022

Abstract

Educational data mining is an emerging interdisciplinary research area involving both education and informatics. It has become an imperative research area due to many advantages that educational institutions can achieve. Along these lines, various data mining techniques have been used to improve learning outcomes by exploring large-scale data that come from educational settings. One of the main problems is predicting the future achievements of students before taking final exams, so we can proactively help students achieve better performance and prevent dropouts. Therefore, many efforts have been made to solve the problem of student performance prediction in the context of educational data mining. In this paper, we provide readers with a comprehensive understanding of student performance prediction and compare approximately 260 studies in the last 20 years with respect to i) major factors highly affecting student performance prediction, ii) kinds of data mining techniques including prediction and feature selection algorithms, and iii) frequently used data mining tools. The findings of the comprehensive analysis show that ANN and Random Forest are mostly used data mining algorithms, while WEKA is found as a trending tool for students' performance prediction. Students' academic records and demographic factors are the best attributes to predict performance. The study proves that irrelevant features in the dataset reduce the prediction results and increase model processing time. Therefore, almost half of the studies used feature selection techniques before building prediction models. This study attempts to provide useful and valuable information to researchers interested in advancing educational data mining. The study directs future researchers to achieve highly accurate prediction results in different scenarios using different available inputs or techniques. The study also helps institutions apply data mining techniques to predict and improve student outcomes by providing additional assistance on time.

Keywords Educational data mining · Predictive analysis · Students attributes

Extended author information available on the last page of the article

1 Introduction

Data mining has shown its success in e-commerce and business development, and now its usages in education sector are growing. Data mining detects patterns in data and be able to mine hidden information in the datasets, which leads to more effective decision making. Data mining is used to withdraw useful information from large datasets. Different approaches are used for data analysis, prediction, and classification to find hidden patterns in a huge meaningless datasets (Adekitan & Salau, 2019). This useful information is mined by the state-of-the-art data mining algorithms. A data mining algorithm is a set of steps that are followed to mine useful information and to build classification and prediction models by finding patterns in the datasets. These algorithms are used in the form of techniques or prediction models. A number of data mining techniques, processes or models to mine data, are being used e.g., Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machine, Artificial Neural Network, and Naïve Bayes. Data mining techniques are becoming beneficial in educational areas as well. Data mining can be used by policy makers to identify essential factors in improving of the education quality. It can also help institutions in analyzing students' achievements, requirements, issues, and learning habits (Adekitan & Noma-Osaghae, 2019).

Higher education is considered as the basis for advancement of a society. We indicate that many students' dropout and withdraw their education or retake admission in the same courses every year. If a massive number of students leave their education because of failure, not only students will suffer themselves, but it will also affect educational systems in a negative way. Therefore, it is necessary to have a system that can detect students' who are going to dropout in their final examinations to minimize failure rates (Chui et al., 2020). No education system is successful if it is not evaluated continuously. In order to improve institutional results and to ensure that all students graduate on time, it is necessary to find out obstacles in the path of students' success. It is very difficult for teachers dealing with many students, to mine their data and detect students' weak areas, but data mining makes it very easy and interesting task without teachers' direct involvement. Data mining techniques help in mining large amount of data in educational sectors to improve teaching and learning processes, called educational data mining. Educational Data Mining (EDM) techniques refer to the methods or algorithms used for mining educational datasets. In EDM, it is essential to extract the required information from huge educational datasets. This information can be used by higher educational authorities to improve policies, by institutions to check and balance teachers'/students' issues and by students to improve their results (Burgos et al., 2018). EDM can be defined as using data mining techniques to immense educational datasets for solving different educational issues. EDM processes involve gathering data, applying models on that data to describe patterns or to mine useful information concerning educational institutes or students. EDM can be used to understand students' learning behaviors and interests to better design teaching strategies that will improve their performance and minimize dropout rates. Educational institutions are storing a huge amount of data every year in their databases. This massive data can be transformed into useful information to help different stakeholders in decision making processes (Kabakchieva, 2013). Since this information, students can detect their weak areas in different courses, teachers can improve their teaching strategies, and administrations can better manage different resources effectively for the benefit of their institution. Data mining makes all these tasks very easy on the behalf of previous experiences and patterns found in the data (Fernandes et al., 2019).

One of the critical tasks in EDM is students' future exam performance prediction. There are several studies published up to the date that used data mining techniques for predicting the students' exam performance. The main goal of these studies is to classify the entire students into two classes, i.e., "pass" or "fail". Students' performance predictions can be conducted by using supervised data mining techniques. In supervised data mining, a mathematical model is built from dataset that describes inputs as well as the desired outputs. It is significant to predict results before many students are dropped out from a specific course. Predicting students' performance is necessary for the institutes to find out weak corners of different courses. This prediction is useful to take an early action by improving learning processes of those students who have high risk of failure in the course.

Some survey studies are published till date that explore research works performed in educational data mining. A survey paper attempts to explore the determinants of students' dropout, in order to benefit future research by highlighting most significant socio-economic features. The study concludes that a mix of individual, economic and educational features affects students' academic outcomes (Aina et al., 2021). Another study focused on supervised data mining algorithms widely used for students' performance prediction (Sen et al., 2020). Two survey studies explored research work to analyze prediction models and students' factors that influence prediction results (Batool et al., 2021; Khan & Ghosh, 2021; Namoun & Alshanqiti, 2021; Qian et al., 2022; Upadhyay et al., 2021). In previous survey studies, we found that, to the best of our knowledge, no survey study tried to explore all factors that may influence students' performance prediction results. The main strength of our survey study is that it summarizes the research work of last two decades with a total of 269 studies and tries to cover all factors that may improve students' exam performance prediction results, called students' attributes. By exploring latest survey papers, another prominent research gap is identified i.e., to the best of our knowledge, no survey paper explored feature selection.

This survey paper analyzed the research work performed in the last two decades by comparing state-of-the-art data mining techniques, data mining tools, and input attributes used for results prediction. The first objective of this paper is to seek for the best prediction techniques. Different prediction techniques are compared in terms of prediction accuracy to find out highly accurate prediction method. The second objective is to identify students' attributes that lead to most accurate prediction results as compared to others. As the third objective, this paper compares and identifies the mostly used data mining tool for prediction process. The presented survey paper synthesizes the machine learning models and tools applied in education to predict student performance. The presented study may help educational institutions to design and deploy a prediction model in their academic sections using available tools and students' attributes. This study may enhance learning management systems (LMS) in virtual learning institutes to predict and prevent students' dropout in online courses. The presented survey enhances the results of previous review studies in order to cover all factors essential in future research work.

We conducted a systematic literature review to answer the research questions:

1) Which data mining algorithm is mostly used in last two decades?
2) Which students' attributes are highly correlated to their exam performance?
3) Which data mining tool is mostly used and why?
4) What is the role of feature selection in students' performance prediction?

This paper answers these questions by comprehensively exploring the latest work and trends in educational data mining. It also focuses on the main aims of all research papers, i.e., to predict students' exam scores or to classify students into pass/fail categories. This study aims to explore the right time for predicting final exam results, too. This study also presents the role of feature selection in predicting students' results. These all factors will direct future research to achieve highly accurate prediction results in different scenarios and using different available inputs or tools.

The remainder of the paper is organized as follows. In Section II presents an overview of survey paper and its contributions. Section III presents a summary of data mining techniques, section IV describes students' attributes and their impact on academic performance, section V gave a comparison of data mining tools used in students' performance prediction. While section VI gives results of presented survey paper and the whole study is concluded in section VII. At last, section VIII and IX presents' limitations of the study and future research work respectively.

2 Method

This section presents the proposed research methodology adopted to conduct the survey. We explore research papers published in the last two decades to answer the research questions mentioned above. We used Google scholar, IEEE Explore, Web of Science, Elsevier and DBLP to find research articles of well-known and impact factor journals, conferences, and thesis published till 2021. This survey paper focused on traditional classroom learning as well as e-learning platforms. The phrases used for searching research articles include "students' performance prediction", "exam score prediction", "educational data mining", "students' academic performance", "students' final exam prediction", "CGPA prediction", "machine learning in predicting students' grades". Using these phrases, a total of 312 research papers were identified, and we stored them in a database. After reading full articles, only 269 research articles were included in this survey because they were focused on the supervised learning techniques. That is, the remaining 43 papers dealt with unsupervised learning techniques. This study focused on classification techniques only, while regression, clustering, association rules and feature optimization methods are

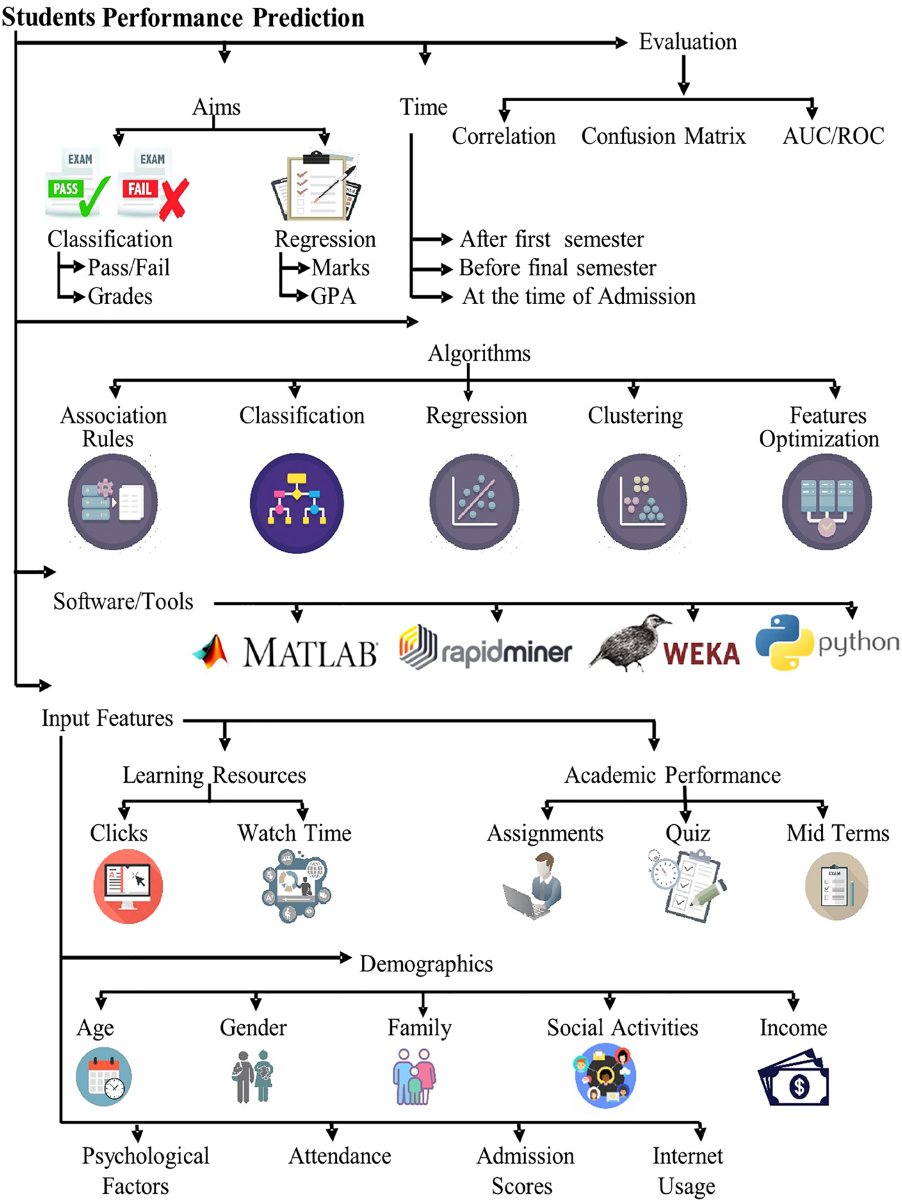

Fig. 1 Survey paper taxonomy

used in few studies to improve the classification results. After literature selection, we presented and summarized the findings of selected articles by comparing and calculating the results. RapidMiner is used to provide a comprehensive analysis of research and the final outcomes are presented in graphs.

Figure 1 presents the directions of this research study. The figure shows that there are mainly two evaluation aims of students' performance prediction i.e., classification and regression. It is presented that research studies evaluate students' performance mainly at the time of admission, mid of academic session and right before final examinations. Data mining algorithms used in studies are presented in the figure, named as association rules, classification, regression, clustering and feature optimization. The four mostly used tools i.e., WEKA, RapidMiner, Python and MATLAB are mentioned in the figure. At last, input features are categorized. The figure presents that main input features are learning resources, academic performance, demographics, psychological factors, attendance, admission scores and internet usage.

3 Data Mining Techniques

There are a number of techniques used for data mining, classification and prediction of the final outcomes. In data mining, classification methods are used for prediction where a classification model distributes a dataset into several classes. Classification process can be divided into two steps. First, based on training data a classifier is generated. Second, this classifier is used to label new data items with unknown classes (Asif et al., 2017). The aim of building a classifier is to make predictions about future data with relevant characteristics by distributing data into predefined classes. In prediction process, different data mining techniques can be adapted to classify students in multiple classes based on their performance, e.g., "pass" or "fail". This section gives a brief overview of different classification techniques used in previous research papers for students' performance prediction.

3.1 A. Decision Tree

One of the mostly used data mining technique for EDM is Decision Tree (DT). Decision Tree is a tree-like graph based on a set of conditions. A set of features are used as input and class labels are the output of Decision Tree. A root node is placed on the top which generates a set of different branches. Each branch describes a condition which is further connected with the next node.

Decision Tree continues this process till it reaches the leaf node. These leaf nodes are labeled as classes or decisions (Tomasevic et al., 2020). Decision Tree follows an IF–THEN algorithm. Decision Tree model is simplest technique and thus it is very easy to understand it's working. Figure 2 describes a simple Decision Tree model which predicts students' results on the basis of some conditions. A study analyzes students' factors before admission and during current semester to predict their semester examination results. Decision Tree is used to build prediction model and study shows 87.14% accurate results (Yathongchai, 2003). A research study applied Genetic algorithm to fine-tune students' score prediction tree (Kalles & Pierrakeas, 2006). Another study (Hsu et al., 2003) used Apriori algorithm to obtain significant factors in predicting students' performance and then applied genetic algorithm for calculating fitness function of variables. A study analyzes students' factors before admission and during current semester to predict their semester examination results. Decision Tree is used to build prediction model and study shows 87.14% accurate results (Yathongchai,

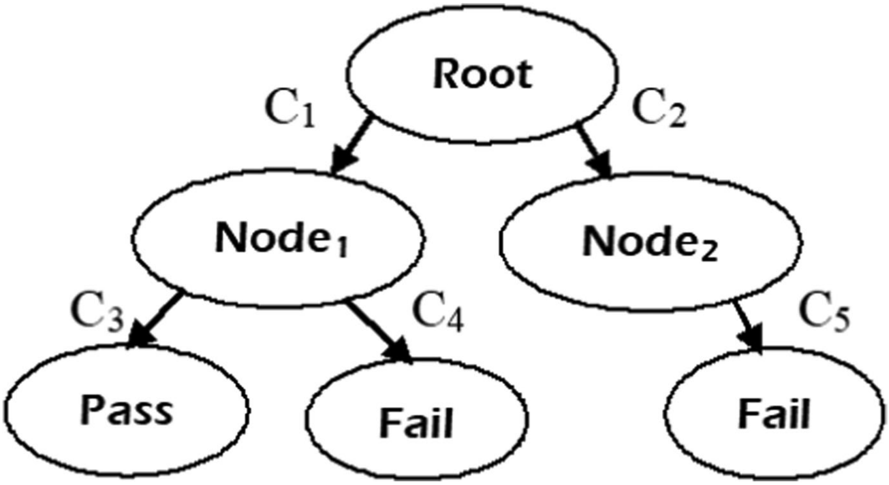

Fig. 2 Decision Tree (DT)

2003). A research study applied Genetic algorithm to fine-tune students' score prediction tree (Kalles & Pierrakeas, 2006). Another study (Hsu et al., 2003) used Apriori algorithm to obtain significant factors in predicting students' performance and then applied genetic algorithm for calculating fitness function of variables.

The study shows that factor analysis positively impacts classification tree results. Generalized sequential pattern mining is used (Patil & Mane, 2014) to draw patterns for predicting students' academic performance. It is proved that using significant features generates more accurate prediction results. Another study proved that Fuzzy genetic algorithm improves the prediction results (Hamsa et al., 2016). Another study used association rule mining to draw a significant correlation between students' admission data and their academic performance (Rojanavasu, 2019). Decision Tree is used to generate prediction rules (Al-Radaideh et al., 2006; Ogor, 2007) and found that students' gender, education funding and CGPA in previous semesters highly influence their final grades. A Decision Tree based early warning system is developed to predict students that are more prone to dropout. Students and teachers are then informed by email and asked to pay more attention in order to improve students' results (Hu et al., 2014). Decision Tree outperformed Neural Network (Herzog, 2006), Naïve Bayes (Nghe et al., 2007) in estimating students' degree completion time and their grades in final examination.

Information Gain and Gain Ratio are used explore correlation between students' factors and their academic performance. It is found that students' study time, age and parents' education highly influence students' results (Osmanbegović et al., 2014). Other studies show that students attendance (Upadhyay & Gautam, 2016) and courses (Altujjar et al., 2016) in current semester are most significant prediction features. An improved Decision Tree is proposed using Information Gain and Entropy. The partition and nodes of Decision Tree are selected with attributes having higher Information Gain. The proposed method is repeated until the best results are obtained (i.e., accuracy = 97.50%)

(Sivakumar et al., 2016; Sivakumar & Selvaraj, 2018). Research studies show that ensemble model gave significant improvements in the prediction accuracy (Jha et al., 2019; Livieris et al., 2018; Pandey & Taruna, 2016).

Students' behavior features and social activities have significant impact of their academic performance, and the research studies recommends using students' cognitive as well as personal, economic and social attributes to predict their exam performance (Aman et al., 2019; Amrieh et al., 2016; Kiu, 2018; Zhao et al., 2020a). Therefore, random features may reduce the prediction accuracy. Dataset preprocessing and significant feature selection enhance the prediction results (Al-Obeidat et al., 2018; Oyefolahan et al., 2018; Wong & Senthil, 2018). An improved ID3 algorithm is proposed comprised of two steps i.e., entropy-based feature selection and prediction model construction. The proposed model proves that eliminating random features enhance model prediction results (Patil et al., 2018, Santoso, 2020).

Several research studies applied Decision Tree for students' performance prediction (Adebayo & Chaubey, 2019; Adhatrao et al., 2013; Akinrotimi et al., 2018; Asif et al., 2017; Banu & Manjupargavi, 2021; Baradwaj & Pal, 2012; Bresfelean, 2007; Buenaño-Fernández et al., 2019; Dey 2020; Figueroa-Cañas & Sancho-Vinuesa, 2020; Hasan, 2019; Hasan et al., 2019; Hew et al., 2020; Kabra & Bichkar, 2011; Kabakchieva, 2013; Kovacic, 2010; Liang et al., 2016; Mikroskil, 2019; Moseley & Mead, 2008; Nandeshwar & Chaudhari, 2009; Patacsil, 2020; Puarungroj et al., 2018; Ramaswami & Bhaskaran, 2010; Sawant et al., 2019; Vivek Raj & Manivannan, 2020; Yadav & Pal, 2012; Zhang & Wu, 2019b). A number of research studies performed a comparison of Decision Tree, Neural Network, Naïve Bayes (Evwiekpaefe et al., 2014), Random Forest (Mishra & Kumawat, 2018; Salal et al., 2019), Lazy Learner (IBK) (Ilic et al., 2016; Meghji et al., 2019; Pandey & Taruna, 2014), KNN (Anuradha & Velmurugan, 2015; Poudyal et al., 2020), Gradient Boosting (Howard et al., 2018), SVM (Anoopkumar & Rahman, 2018; Francis & Babu, 2019), Logistic Regression (Perez et al., 2018; Salal et al., 2019; Kemper et al., 2020), Multilayer Perceptron (MLP) (Freitas et al., 2020) and Sequential Minimal Optimization (Acharya & Sinha, 2014) and proved that Decision Tree is most effective prediction model. In last decades, several Decision Tree algorithms have been used as classification and prediction models. In (Hamoud, 2016; Hamoud et al., 2018), J48, Random Tree, Hoeffding and Rep tree are found as best Decision Trees for students' academic performance prediction. Different Decision Tree models are compared namely CART, CHAID, C4.5 and ID3 and proved that JRip (Walia et al., 2020), CART (Saa, 2016; Wong & Yip, 2020) and C4.5 (Saheed et al., 2018) gave the best prediction results. Another study shows that JRip prediction model outperformed as compared to other Decision Tree classifiers (Pattanaphanchai et al., 2019). Table 1 presents a summary of research studies implementing Decision Tree for students' performance prediction.

Table 1 Decision tree for students' performance prediction

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2003	University (BRU) Dataset (2008–2009)	-	Pre-Admission Grades, School Details, Loan, Dropout Semester & Cause	Accuracy = 87%	(Yathongchai, 2003)
2003	UCI Repository Dataset (2003)	Association Rules & Genetic Algorithm	Department, Courses Type, Credit Hours & Academic Performance	Accuracy = 80%	(Hsu et al., 2003)
2006	Hellenic Open University Dataset (2006)	-	Demographics & Academic Performance	Accuracy = 92%	(Kalles & Pierrakeas, 2006)
2006	University Dataset (2006)	Gain Ratio	Demographics, Courses, Study Type, Funding, Instructor's Details & Grades	Accuracy = 38%	(Al-Radaideh et al., 2006)
2006	University Dataset (2006)	-	Demographics, Residence, Financial, Parents' Income & Academic Performance	Accuracy = 93%	(Herzog, 2006)
2007	University Dataset (2003–2005)	Information Gain	Demographics, Religion, Study Groups, TOEFL, Funds/Scholarships & Admission Scores	Accuracy = 94%	(Nghe et al., 2007)
2007	University Dataset (2005–2006)	-	Demographics & Assessments' Data	Accuracy = 97.3%	(Ogor, 2007)
2007	High School Dataset (2007)	-	Pre-Admission Education, Learning Resources, Research, Job, Scholarships, Parental Support & Final Degree	Accuracy = 88%	(Bresfelean, 2007)
2008	Nursing School Dataset (2008)	CHAID	Age, Gender, Entry Test Grades, Branch, Student's Performance & Attendance	Accuracy = 94%	(Moseley & Mead, 2008)
2009	WVU's Data Warehouse (1999–2006)	Wrapper & Information Gain	Demographics, Admission & Academics Factors	Accuracy = 84%	(Nandeshwar & Chaudhari, 2009)

Table 1 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2010	Self-Designed Questionnaire (2010)	CHAID Feature Selection	Health, Family Details, School Details, Learning Activities	Accuracy = 44.69%	(Ramaswami & Bhaskaran, 2010)
2010	Open Polytechnic Student Management System (2010)	-	Demographics, School Qualification, Courses & Contact Details	Accuracy = 59.4%	(Kovacic, 2010)
2011	VBS Purvanchal University, India (2011)	-	-	-	(Baradwaj & Pal, 2012)
2011	University Dataset (India, 2011)	-	Demographic Data, SSC, HSC Exam Marks, Address & Contact Number	TP Rate = 0.907	(Kabra & Bichkar, 2011)
2012	VBS Purvanchal University, India (2010)		Basic Information, Past Performance, Address & Contact Number	Accuracy = 67.7%	(Yadav & Pal, 2012)
2013	University Admissions & Results Database (2013)	-	Admission Data & Exam Grades	Accuracy = 67%	(Kabakchieva, 2013)
2013	College Database (2013)	Information Gain	SSC, HSSC & Entrance Scores	Accuracy = 75.14%	(Adhatrao et al., 2013)
2014	College Database (2014)	Generalized Sequential Pattern Mining	Personal/Family Background & Academic Factors	Accuracy = 91.11%	(Patil & Mane, 2014)
2014	College Database (2014)	Gain Ratio	Age, Location, Study Gap & Previous Grades	Accuracy = 100%	(Pandey & Taruna, 2014)
2014	National University Database (2009–2010)	Gain Ratio & Gini Index	Login, Course Materials Usage, Assignments & Discussion Status	Accuracy = 98%	(Hu et al., 2014)
2014	University Database (2014)	Correlation, Chi-Square Based Feature Evaluation & Information Gain	Basic Information, Attendance & Mid-Term Scores	F-Measure = 79%	(Acharya & Sinha, 2014)

Table 1 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2014	Secondary Schools Data (2011–2012)	Gain Ratio & Info Gain	Basic Information, Parents' Income, Study Duration, Internet Access, Travel Time to School etc	Accuracy = 74%	(Osmanbegović et al., 2014)
2015	College Database (2015)	-	Scores in Previous Classes, Family Factors, Living Location & Attendance	Accuracy = 70%	(Anuradha & Velmurugan, 2015)
2016	Educational Institute Dataset (2016)	Fuzzy Set Theory	Previous Class Percentage, Attendance, Sports Interest & Parents Economic Status	Accuracy = 99%	(Upadhyay & Gautam, 2016)
2016	University Dataset (2016)	Entropy and Information Gain	Family Education & Profession, Campus Environment, Syllabus, Course Satisfaction & Extra-Curricular Activities	Accuracy = 97.5%	(Sivakumar et al., 2016)
2016	School Database (2013–2015)	-	Attendance in Classroom, Lab & Test Scores	Accuracy = 97.2%	(Ilic et al., 2016)
2016	King Saud University (2013–2014)	-	College Exam Percentage, University GPA, GAT Scores, Courses & Assessments Marks	Accuracy = 80%	(Altujjar et al., 2016)
2016	Kalboard 360 (2016)	Wrapper & Filter Based Feature Ranking	Demographics, Academic Performance & Behavioral Features	Accuracy = 79%	(Amrieh et al., 2016)
2016	Massive Open Online Courses (MOOC) Dataset (2015)	-	Students' Web Behavior	Accuracy = 88%	(Liang et al., 2016)
2016	UCI Repository Dataset (2014)	-	Personal Information, Study Time, School & Family Details, Attendance & Alcohol Consumption	Accuracy = 91.9%	(Hamoud, 2016)

Table 1 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2016	University Dataset (2016)	-	Admission Data, Attendance, Assignments & Exam Scores	Accuracy = 80%	(Hamsa et al., 2016)
2016	Online Survey (2016)	Naïve Bayes	Nationality, Gender, Family Details, Exam Scores, Transport & Friends	Accuracy = 40%	(Saa, 2016)
2016	College Database (2016)	-	Academic Performance	Accuracy = 98%	(Pandey & Taruna, 2016)
2017	Public Sector Engineering University, Pakistan (2007–2009)	Gini Index & Information Gain	Admission Data & Exam Scores	Accuracy = 83.6%	(Asif et al., 2017)
2017	UCI Repository Dataset (2014)	PROAFTN	Personal & Family Factors, Social Activities, Interest in Studies & Attendance	Accuracy = 82%	(Al-Obeidat et al., 2018)
2017	University Database (2017)	LSTM Based Feature Selection	Students' Behavioral Patterns	Accuracy = 87%	(Zhao et al., 2020a)
2018	University Dataset (2015–2016)	-	Basic Information, Assessment Grades & LMS Usage	MAE = 6.5%	(Howard et al., 2018)
2018	University Dataset (2013–2015)	Pearson's Correlation Coefficient	Gender, Faculty, Blood Type & GPA	Accuracy = 81.62%	(Puarungroj et al., 2018)
2018	Nigerian University (2013–2014)	-	Personal Information, Parents' Occupation, Courses, Mode & Year of Admission	Accuracy = 98%	(Saheed et al., 2018)
2018	University Dataset (2015–2017)	Information Gain	Basic Information, SSC Marks, State, Hostel, Sports, Registration & Academic Scores	Accuracy = 93.3%	(Oyefolahan et al., 2018)
2018	University Dataset (2018)	Information Gain	Academic Marks & Attendance	Accuracy = 97.7%	(Sivakumar & Selvaraj, 2018)
2018	University Dataset (2004–2010)	PCA	Admission Details, Demographics, Financial Aids & Exam Scores	AUC = 94%	(Perez et al., 2018)

Table 1 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2018	University Dataset (2018)	Information Gain	Name, Gender, Caste, Admission Year, Type & Marks	Accuracy = 74%	(Patil et al., 2018)
2018	University Dataset (2018)	Info Gain, Gain Ratio & Entropy	Personal, Residence, Academic Details & Attendance	Accuracy = 91.3%	(Anoopkumar & Rahman, 2018)
2018	College Dataset (2018)	Apriori Algorithm	Basic Information & Courses	Accuracy = 97.43%	(DEY 2020)
2018	University Dataset (2018)	-	Gender, Parents Education & Profession, Residence, Medium of Instruction, Degree Type, Academic Duration & Scores	Accuracy = 94%	(Mishra & Kumawat, 2018)
2018	UCI Repository Dataset (2014)	-	Parents' Education & Job, Family Size, Financial Support, Study Time, Extracurricular/Social Activities, Health & Failures	Precision = 0.924	(Kiu, 2018)
2018	Online Questionnaire (2018)	Correlation Based Attribute Evaluation	Basic Information, Parents' Working Details, Emotional Health, Interest Level in Studies, Drugs Consumption, Financial Resources & Confidence	Precision = 0.629	(Hamoud et al., 2018)
2018	MCA Students Dataset (2018)	Chi-Square, Oner Attribute Evaluation	Personal, Family & Academic Details	Accuracy = 90.6%	(Wong & Senthil, 2018)
2018	School Students' Dataset (2012–2016)	-	Students' Grades	Accuracy = 87.15%	(Livieris et al., 2018)
2018	Kwara State University, Nigeria (2013–2015)	Information Gain	Parents' Education/Profession & Students' Grades	Accuracy = 92%	(Akinrotimi et al., 2018)
2019	University Dataset (2012–2014)	-	Demographics & Academic Performance	Accuracy = 78%	(Mikroskil, 2019)

Table 1 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2019	School Assessment Results Data (2019)	-	Quiz Results	-	(Adebayo & Chaubey, 2019)
2019	UCI Repository Dataset (2014)	Correlation, Gain Ratio & Information Gain	Demographics, Social Activities, Academic Grades & Attendance	Accuracy = 76.7%	(Salal et al., 2019)
2019	University Dataset (2019)	-	Classroom Behaviors & Grades	Accuracy = 88.23%	(Meghji et al., 2019)
2019	University Dataset (2016–2018)	-	Assessments' Grades	TP = 93%	(Buenaño-Fernández et al., 2019)
2019	University Dataset (2019)	-	Assessment Marks & Attendance	Accuracy = 94%	(Hasan et al., 2019)
2019	University Dataset (2016–2018)	-	Demographics, Attendance & Grades	Accuracy = 78%	(Hasan, 2019)
2019	College/University Dataset (2019)	Decision Tree	Academic Performance	Accuracy = 81%	(Sawant et al., 2019)
2019	University Dataset (2019)	-	Demographics, Academic & Behavioral Attributes	Accuracy = 75%	(Francis & Babu, 2019)
2019	University Dataset (2017–2018)	Information Gain & Gain Ratio	Basic Information, Lab Scores, Number of Questions Solved, Lecture Watch Time & Forum Discussions	Accuracy = 73%	(Zhang & Wu, 2019)
2019	Open University Learning Analytics Dataset (OULAD, 2015)	-	Demographics, Sum of Clicks & Assessments' Marks	AUC = 0.93	(Jha et al., 2019)
2019	MOOC (2015)	-	Courses Schedule, Duration & Time Spent	F1-Score = 88%	(Hew et al., 2020)
2019	Admission Dataset (2016–2017)	Information Gain	Academic Year, Department, School & Selected Courses	Accuracy = 73.58%	(Rojanavasu, 2019)
2019	Prince Of Songkla University Dataset (2013–2017)	-	Admission Scores, Demographics &	Accuracy = 77%	(Pattanaphanchai et al., 2019)

Table 1 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2019	University in Pakistan Dataset (2019)	-	Demographic, Social & Academic Details	F-Measure = 0.835	(Aman et al., 2019)
2019	UCI Repository Dataset (2014)	Information Gain & Gain Ratio	Demographics, Social Activities, Health, Attendance & Grades	Accuracy = 76.7%	(Salal et al., 2019)
2020	University Dataset (2018–2019)	-	Academic Details	F-Measure = 76%	(Figueroa-Cañas & Sancho-Vinuesa, 2020)
2020	Chinese University of Hong Kong Dataset (2020)	Gain Ratio	Attendance, Assignments, Quizzes & Presentation Scores	Accuracy = 92%	(Wong & Yip, 2020)
2020	University Dataset (2020)	-	Gender, Race, High School & Income	Accuracy = 99%	(Freitas et al., 2020)
2020	University Dataset (2007–2012)	-	Health, Family Issues, Examinations & Grades	Accuracy = 95%	(Kemper et al., 2020)
2020	School Dataset (2012–2016)	-	Registered Courses & Grades	Accuracy = 70%	(Patacsil, 2020)
2020	Kwik Kian Gie School of Business Dataset (2020)	Entropy and Information Gain	GPA & Attendance	Accuracy = 95.85%	(Santoso, 2020)
2020	LMS Database (2020)	-	-	Accuracy = 71%	(Banu & Manjupargavi, 2021)
2020	University Dataset (2020)	-	Personal Information, Family Education, Business, Years of Study, Number of Siblings & Accommodation	Accuracy = 72.44%	(Vivek Raj & Manivannan, 2020)
2020	Nigerian Defense Academy (2020)	Wrapper Subset Evaluation	Demographic & Academic	Accuracy = 69.7%	(Evwiekpaefe et al., 2014)
2020	UCI Repository (2014)	CorrelationAttributeEval	Demographic, Attendance, Social Activities & Academic Features	Accuracy = 85.31%	(Walia et al., 2020)
2020	OULAD (2015)	PCA	Demographic, VLE Clicks and Assessments' Scores	Accuracy = 99.1%	(Poudyal et al., 2020)

3.2 B. Naïve Bayes

Naïve Bayes (NB) is another classification algorithm based on the Bayesian theorem. It is called Naive as this technique assumes that there is no hidden relationship between data attributes that can affect prediction results. It calculates the probability of belonging to a specific class. The class which obtains highest probability is considered as the class of that data (Tomasevic et al., 2020). Below equation shows the Bayesian formula which calculates probability of class A with the association of class B.

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)*P(A)}}{\mathrm{P(B)}} \tag{1}$$

Naïve Bayes algorithm has the potential to predict students' academic performance (Bekele & McPherson, 2011; Bekele & Menzel, 2005). A Bayesian model is proposed to classify students into different classes based on their academic performance (Ramaswami & Rathinasabapathy, 2012). Naïve Bayes is also compared to KNN for the selection of most accurate prediction model. The study shows that Naïve Bayes achieved more accuracy i.e. 93.6% when both models were constructed using demographic features only (Amra & Maghari, 2017). Another study used Naïve Bayes algorithm to forecast students' grades in their final exam. Different semester activities e.g. students' assignments, previous grades, attendance and lab assessments are proved to be very useful features for prediction of final exam grades (Shaziya et al., 2015). Naïve Bayes and SVM prediction model are built and compared, where Naïve Bayes gave better results i.e. 92% (Tripathi et al., 2019) and 63.5% (Kaur & Bathla, 2018). Distance learning makes it more challenging for tutors to interact with individual students, identify their weak areas and to predict students' academic performance. A study implemented Naïve Bayes algorithm to provide a supporting tool for teachers which predict students' final exam performance in distance learning environment (Kotsiantis et al., 2002). Students' demographic variables including parents' qualification, jobs, living status, income, students' eating habits are used to develop Naïve Bayes prediction model. The prediction model found that students' scores in secondary school, living status and medium of teaching are highly correlated to their grades in college examinations (Bhardwaj & Pal, 2012).

A comparison of different classification algorithms namely Decision Tree, Random Forest, Naïve Bayes, MLP, KNN and Logistic Regression is performed, and results show that Naïve Bayes gave best prediction results as compared to other techniques (Koutina & Kermanidis, 2011; Romero et al., 2013; Barbosa Manhães et al., 2015; Marbouti et al., 2016; Yaacob et al., 2019; Ahmed et al., 2020a). Naïve Bayes, MLP and J48 algorithms are used for students' exam performance prediction based on their previous academic performance. The study shows that Naïve Bayes gave best results i.e. accuracy = 76.65% (Osmanbegovic & Suljic, 2012). Naïve Bayes and Decision Tree classification algorithms are compared, and it is found that Naïve Bayes outperform in predicting students' final semester marks. Students' demographic and academic attributes are preprocessed to improve classifier's accuracy

(Kaur & Singh, 2016; Khasanah, 2017; Mueen, 2016; Wati et al., 2017). Similarly, students' admission test scores (Harvey & Kumar, 2019) and final exam results (Kumar et al., 2019) are predicted using Naïve Bayes and Decision Tree prediction models. The study shows that Naïve Bayes gave higher accuracy of 71% and 85% respectively. A web based Naïve Bayes classifier is developed to store students' data, retrieve useful information and to predict their final exam success rate. Such a prediction model is found very useful for institutions to maintain their success graph and to provide relevant assistance to students and teachers (Devasia et al., 2016).

Most of the research studies focus on students' family background and previous academic performance to predict their future exam scores. However, students' personality is also a contributing factor which highly affects their educational interests. This study focused on time management, leadership, self-reflection, social support, study preference and future to predict how they are going to perform in their future exams. It is found that non-cognitive features support cognitive features to increase accuracy of Bayesian prediction model (Sultana et al., 2017). Similarly, another study focused on neglected features namely family expenditures, income, and family assets to explore their impact on students' academic performance. Using SVM and Naïve Bayes prediction models, the study found that students' performance is highly correlated to their family utility bills and expenses on education. A decrease in other expenditures may increase the opportunities to complete their higher education (Daud et al., 2017). A common objective of almost all mentioned studies is to build an early prediction model so that students can be prevented from dropout. A weekly approach is used to predict students' final exam scores after each week of their admission before the final exams. The results show that adding more events to the dataset may increase prediction accuracy i.e. 73.5% after week 1 and 77.7% after week 16 (Akçapınar et al., 2019). Feature optimization is used to remove irrelevant features from the dataset. Forward Selection (Saifudin & Desyani, 2020), PCA (Borges et al., 2018) and Wrapper (Usman et al., 2020) feature selection techniques used with Naïve Bayes model enhanced students' performance prediction results and also reduced the time required for model construction. Table 2 presents a summary of research studies implementing Naïve Bayes for students' performance prediction.

3.3 C. Artificial Neural Network

Artificial Neural Network (ANN) is a well-known classification technique used to solve data mining problems. The concept of ANN is based on the biological neural network. ANN model is divided into three layers, input layer is used to take input data, hidden layer consists of a set of neurons that process data and output layer gives final classes of the data (Amazona & Hernandez, 2019). Input neurons are connected to the next neurons in hidden layer, in order to transmit a signal for processing. These hidden neurons process signal and forward it to the next connected neurons, until the signal reaches to output layer. Branches are used to connect neurons with each other. These branches are assigned with some weights to set the strength of the signal (Tomasevic et al., 2020). Figure 3 describes the working of a basic ANN model which predicted binary classes, i.e., pass or fail.

Table 2 Naive Bayes for students' performance prediction

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2003	University Dataset (2003)	Information Gain	Demographics & Academic Performance	Accuracy = 77%	(Kotsiantis et al., 2002)
2005	High School Dataset (2005)	Random Search Algorithm	Gender, Learning Behaviors, Confidence, English & Mathematics Performance	Accuracy = 78%	(Bekele & Menzel, 2005)
2011	Secondary School Dataset (2011)	-	Academic Achievement, Motivation, Personality, Family & Educational Background	Accuracy = 78.4%	(Bekele & McPherson, 2011)
2011	University Database (2011)	Filtered Feature Selection Technique	Students' Background, Location, Parents' Qualification & Occupations	P = 0.8642	(Bhardwaj & Pal, 2012)
2012	Questionnaire Conducted In Faculty Of Economics In Tuzla (2010–2011)	-	Students' Basic Information, GPA, Scholarships & Course Materials	Accuracy = 76.65%	(Osmanbegovic & Suljic, 2012)
2012	School Students' Data (2009)	Consistency Subset Evaluation, Chi-Square & Information Gain	Demographic Details, Family Details, Socio-Economic Details & Previous Academic Performance	Accuracy = 84.8%	(Ramaswami & Rathinasabapathy, 2012)
2013	Moodle Online Discussion Forums (2013)	Filtering Based Attribute Selection	Number of Posts, Time Spent Online & Evaluation Scores by Teacher	Accuracy = 90.3%	(Romero et al., 2013)
2015	AMSS Database (2010)	-	Enrollment Year, Program, Grading Criteria, Credit Hours, Obtained Grades, Passing Grades & Attendance	Accuracy = 86.57%	(Barbosa Manhães et al., 2015)
2016	U.S. University Dataset (2013–2014)	Pearson Correlation	Academic Performance	Accuracy = 94.9%	(Marbouti et al., 2016)

Table 2 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2016	-	-	Gender, Hometown, Previous Grades, Attendance, Sports & Family Factors	Accuracy = 63.59%	(Kaur & Singh, 2016)
2016	University Dataset (2014–2015)	Attribute Ranker	Demographics, Academic Performance & Web Login Details	Accuracy = 86%	(Mueen et al., 2016)
2016	College Database (2013–2016)	-	Personal, Family Details, Exam Scores, Study Time, Reading Habits, Interest in Higher Studies & Access to Mobile/Internet	-	(Devasia et al., 2016)
2017	NUST Pakistan, Dataset (2017)	-	Demographics, Cognitive & Non-Cognitive Features	Accuracy = 84%	(Sultana et al., 2017)
2017	Universitas Islam Indonesia's Information System (UNISYS, 2017)	Correlation, Gain Ratio, Info Gain, Relief & Symmetrical Uncertainty	Gender, Family Factors, High School Type & Grades, Attendance & GPA	Accuracy = 98.08%	(Khasanah, 2017)
2017	University Datasets (2004–2011)	Information Gain & Gain Ratio	Personal Information, Family Income, Assets & Expenditures	F1-Score = 86.7%	(Daud et al., 2017)
2017	University Database (2010–2012)	-	Age, Gender, Programs, Graduation Time & GPA	Accuracy = 76.79%	(Wati et al., 2017)
2017	Secondary Schools Dataset (2017)	-	Enrollment Year, Contact, Basic Information & Marital Status	Accuracy = 93.6%	(Amra & Maghari, 2017)
2018	MCA students' dataset (2018)	CFS Subset Evaluation & Greedy Stepwise Search	Academic Performance	F-Measure = 72.7%	(Shaziya et al., 2015)
2018	Moodle Dataset (2018)	-	Learning Resources, Assignments, Quizzes & Personal Information	Precision = 86%	(Helal et al., 2018)

Table 2 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2018	UCI Repository Dataset (2014)	PCA	Parents' Education, Job & Financial Status, Syllabus, Internet, Health, Extra-Curricular Activities, Study Hours & Attendance	F1-Score = 99%	(Borges et al., 2018)
2018	Kaggle Dataset (2015)	-	Demographics, Parents' Education/Occupation, Social Activities, Number of Failures & Attendance	Accuracy = 63.6%	(Kaur & Bathla, 2018)
2019	Questionnaire (2019)	Density Based Clustering	-	Accuracy = 92%	(Tripathi et al., 2019)
2019	Bookroll System (2019)	-	Click-Stream Data	Accuracy = 84%	(Akçapınar et al., 2019)
2019	University Dataset (2013–2016)	Information Gain & Gini Index	Gender, Courses Scores & GPA	Accuracy = 89%	(Yaacob et al., 2019)
2019	School Dataset (2019)	Correlation Among Attributes	Students' Demographics, School Financial Position, Assessments & Results	Accuracy = 71%	(Harvey & Kumar, 2019)
2019	-	-	-	Accuracy = 85%	(Kumar et al., 2019)
2020	University Dataset (2009–2015)	CFS & Wrapper	Academic Performance	Accuracy = 91%	(Usman et al., 2020)
2020	UCI Dataset (2014)	Forward Selection	Demographic, Social Activities, Attendance & Academic Features	Accuracy = 94.43%	(Saifudin & Desyani, 2020)
2020	Ionian University, Corfu (2020)	-	Demographic & Exam Performance	Accuracy = 100%	(Koutina & Kermanidis, 2011)

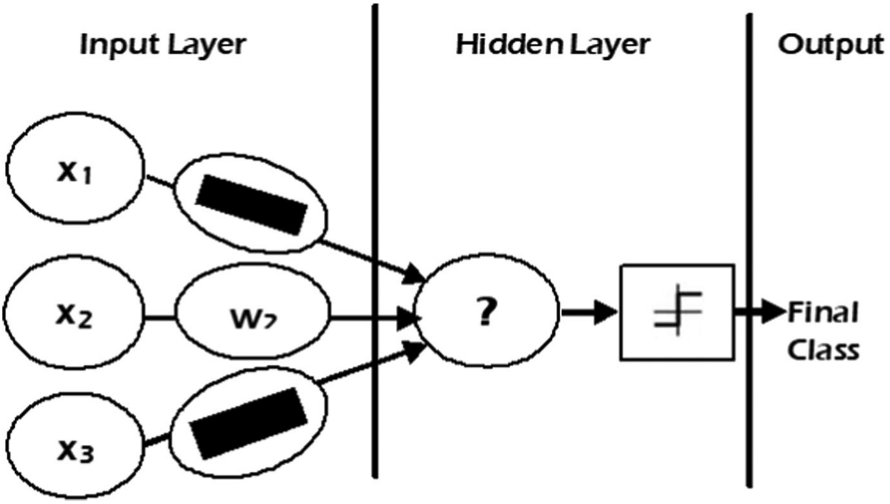

Fig. 3 Artificial Neural Network (ANN)

Several studies present the utilization of ANN for students' academic performance prediction (Calvo-Flores et al., 2006; Karamouzis & Vrettos, 2008; Lykourentzou et al., 2009; Paliwal & Kumar, 2009; Wook et al., 2009; Arsad & Buniyamin, 2013; Agrawal & Mavani, 2015; Whitehill et al., 2017; Altaf et al., 2019; Liu, 2019; Mi, 2019; Raga & Raga, 2019; Sukhbaatar et al., 2019; Umar; 2019; Khazaaleh, 2020; Sood & Saini, 2020).

A comparative analysis of different classification algorithms shows that ANN and Random Forest gave better prediction with more accurate results (Alloghani et al., 2018).

Ensemble methods are used to strengthen the prediction results. ANN model predicts results more accurately when supported by ensemble filtering method (Rahman & Islam, 2017). Another study hybrid wrapper feature selection with four prediction algorithms i.e., Decision Tree, Naïve Bayes, KNN and CNN to enhance the accuracy of individual model. The study shows that CNN model outperformed i.e. accuracy = 95% (Turabieh, 2019). Multi-layer Perceptron (MLP) is also used to predict students' performance (Ahmad & Shahzadi, 2018; Ruby & David, 2015). MLP classifier consists of multiple layers, where each layer has a different function to perform. A predictive model (Olalekan et al., 2020) consisting of two layers: ANN and Naïve Bayes is proposed and found that MLP gave better prediction results as compared to single models. Multi-layer ANN model (Yağci & Çevik, 2019) predicted successful students in different subjects with an accuracy up to 99%. In several studies (Ramesh, 2013; Kaur et al., 2015, Yahaya et al., 2020) MLP gave better results (i.e., accuracy = 72.38% and 75% respectively) as compared to other classification models. A research study collected students' attributes from school's database. These attributes are pre-processed, and sparse auto encoder algorithm is applied to forecast influential factors from random features. Then, the MLP model is trained using influential factors only which gave better prediction results as compared to prediction with all random variables (Guo et al., 2015).

ANN prediction results depend on the number and type of input data on which model is trained. The study shows that students' cognitive and non-cognitive variables have significant importance in their final exam results and its prediction (Lin, 2009). Data pre-processing is a significant step which helps to enhance machine learning algorithms' results. In a study, Synthetic Minority Over-Sampling (SMOTE) is applied on students' dataset and it proves that pre-processing shows a significant increase in prediction accuracy i.e. up to 7% (Jishan et al., 2015). E-learning made it easier for teachers and institutions to record students' interactions, clicks, study-time, durations, assignment submission, learning habits etc. On the other hand, traditional classrooms have limited information related to students cognitive and non-cognitive attributes. A study proposed ANN based prediction model with limited number of students' attributes and achieved 62.5% accuracy (Chanlekha & Niramitranon, 2018). Feature selection is used to find correlation between students' academic features and results. A comparison of ANN prediction model with highly-correlated features and with random features is performed (Hamoud & Humadi, 2019). The study proves that several students' features do not participate or take less part in predicting students' results. Another study examined the contribution of input features to the prediction of output classes. It shows that students' attendance and study duration are the best input variables for ANN based students' results prediction model (Aydoğdu, 2020). PSO is applied before providing input values to the ANN back-propagation model, which increased the prediction accuracy and decreased the number of iterations (Sari & Sunyoto, 2019). A comparative analysis of different supervised learning algorithms and different students' attributes is performed (Tomasevic et al., 2020). The study shows that ANN based on students' assessments' marks and interaction with learning material is the best prediction model. Different data mining algorithms are compared i.e., Decision Tree, Random Forest, Naïve Bayes, KNN, SVM, Logistic Regression and Neural Network and it is found that Neural Network outperformed other algorithms with highest accuracy (Cavazos & Garza, 2017; Vijayalakshmi & Venkatachalapathy, 2019; Bravo et al., 2020; Makombe & Lall, 2020; Mengash, 2020; Waheed et al., 2020). Decision Tree, ANN and Regression algorithms are compared and found that ANN gave best results (Mutanu & Machoka, 2019). In computer programming courses, students' assignments' completion is found to be a significant factor that highly influence their semester results (Qu et al., 2019). Deep Learning model gave an accuracy of 82.5% in predicting students' success in programming courses (Pereira et al., 2020). Neural network is implemented to explore association between students' internet usage and their academic results. The study found that high achievers spent more time on internet, however they have low download and upload volume. It is concluded that students' spent a lot of time using internet but their usage behavior is quite different and can be used to predict their success in exams (Xu et al., 2019). Different research studies have used different statistical software to perform predictive analysis. A study is proposed to build ANN prediction model using two different platforms i.e., SPSS and MATLAB. The study shows that ANN model prediction results are higher than results in SPSS (Çevik & Tabaru-Örnek, 2020). Studies also show that CNN (Karimi et al., 2020; Zong et al., 2020), Deep Belief Network (Sokkhey & Okazaki, 2020a, 2020b, 2020c), Deep Learning (Amazona & Hernandez, 2019; Hussain et al., 2019), LSTM (Su et al., 2018, Li, 2020; Liu et al., 2020) can gave acceptable prediction results. In a study (Karlık & Karlık, 2020), different neural networks namely ANN, MLP, DNN and CNN are used to build students' prediction models and the comparison shows that Fuzzy CNN model gave an accuracy of 92%. Another study proved that Regression Neural Network gave better results as compared to MLP (Iyanda, 2018). A hybrid algorithm based on RNN, gated recurrent unit (GRU) and LSTM is proposed. The experimental results show that model prediction results depend on input parameters, however, the proposed model achieved better accuracy (i.e. up to 80%) as compared individual models (He et al., 2020). Table 3 presents a summary of research studies implementing ANN for students' performance prediction.

3.4 D. Support Vector Machine

Support Vector Machine (SVM) divides the dataset belonging to different classes by using model approach. It plots data items on a 2 or 3-dimensional space and draw a hyper plane between two different classes. The items that fall on one side of hyper plane are considered as belonging to one class. The data item nearest to hyper plane is called vector. A wider hyper plane represents a better separation of data as it clearly presents two different classes. SVM is specifically proposed for binary classification but several algorithms are used under SVM to solve multi class problems (Sen et al., 2020). Being a weight-based method, SVM is used for classification as well as feature extraction. Figure 4 presents the binary classification model using SVM.

Support Vector Machine classification model is used to predict engineering students' final exam score based on their previous exam performance. SVM model is compared with linear regression and multilayer perception to find best prediction model among these classifiers. The results show that SVM gave best results i.e. accuracy = 90.1% (Huang & Fang, 2013) and 50% (Mativo & Huang, 2014). A high number of students drops out in programming courses, which highly affects their final GPA. Information Gain (IG) is used to explore different students' variables and to assign weights to highly correlated features. The highly effective attributes are then used to develop SVM prediction model. The study found that students' mid-term exams are best predictors for their final exam scores (Costa et al., 2017). Multi-level SVM based prediction model is developed which classifies students into five levels, on the basis of their GPA (Asogbon et al., 2016). Similarly, a three-level SVM classifier predicts 90% accurate results. The proposed models may help institutes to place students into different sections and to provide them the required attention (Burman & Som, 2019). A three-step prediction model namely students' entrance in college, after first semester and after second semester is developed. At the first step, students' attributes available at the time of admission are used for prediction, however, students' academic results are added afterward. The proposed study is significant to predict students' performance as early as possible (Gil et al.,

Table 3 Ann for students' performance prediction

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2006	Moodle Dataset (2006)	-	LMS Activities	Accuracy = 80%	(Calvo-Flores et al., 2006)
2008	Waubonsee College Dataset (1997–2002)	-	Demographics, Residence, Major Subjects & Grades	MAE = 0.36	(Karamouzis & Vrettos, 2008)
2009	Moodle E-Learning Course (2007)	-	Test Scores	MAE = 0.63	(Lykourentzou et al., 2009)
2009	University Dataset (2009)	-	Demographic & Academic Factors	Accuracy = 71%	(Lin, 2009)
2009	Indian Schools Dataset (2009)	-	Test Grades, Group Discussion, Interview, Working Experience & Academic Performance	Accuracy = 74%	(Paliwal & Kumar, 2009)
2009	University Dataset (2008–2009)	-	Demographics, Personality, Previous Qualification, Test Scores & Computer Skill	-	(Wook et al., 2009)
2013	University Students' Data, Malaysia (2005–2008)	-	CGPA	MSE = 0.05544	(Arsad & Buniyamin, 2013)
2013	Secondary School Data (2013)	Chi-Squared, Info Gain, Oner & ReliefF	Basic Information, Parents' Economic Status, Tuition, Core & Elective Subjects, Computer & Internet Access Etc	Accuracy = 72.38%	(Ramesh, 2013)
2015	High School Dataset (2015)	Chi-Squared, Info Gain & ReliefF	Basic Information, Attendance, School Information, Computer & Internet Access	Accuracy = 75%	(Kaur et al., 2015)
2015	College Dataset (2012)	Chi Square, Info Gain, Gain Ratio, Correlation & Regression	Student Attendance, Personal & Academic Details	Accuracy = 83.6%	(Ruby & David, 2015)
2015	University Database	Naïve Bayes	Scores, Location & Medium of Instruction	Accuracy = 70.48%	(Agrawal & Mavani, 2015)

Table 3 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2015	100 Junior High Schools Data	-	Demographics, Assessment Scores, Courses & School Details	Accuracy = 86.5%	(Guo et al., 2015)
2015	University Dataset (2015)	-	Academic Scores & Attendance	Accuracy = 75%	(Jishan et al., 2015)
2017	Knowledge Discovery in Databases (KDD) Cup 2015	-	Students' Course Interactions	Accuracy = 97.55%	(Whitehill et al., 2017)
2017	Kalboard 360 (2010)	-	Demographics, Academic, Behavioral Attributes & Parent's Participation in Students' Studies	Accuracy = 84.3%	(Rahman & Islam, 2017)
2018	SETAP Machine Learning Data (2012–2017)	-	Student's Efficiency & Understanding	Accuracy = 65.9%	(Alloghani et al., 2018)
2018	Kasetsart University (2008–2017)	-	Gender, Graduation Year, Admission Type, Department & Grades	Accuracy = 62.5%	(Chanlekha & Niramitranon, 2018)
2019	University of Basra Dataset (2019)	Information Gain, Correlation, SVM & PCA	Department, Family, Academic Performance, Activities, Residence & Self-Esteem	Precision = 87%	(Hamoud & Humadi, 2019)
2019	High School Dataset (2019)	-	Demographics & School Type	Accuracy = 98%	(Yağci & Çevik, 2019)
2019	University Dataset (2017–2018)	-	Gender, Department & LMS Activities	Accuracy = 80.47%	(Aydoğdu, 2020)
2019	University Dataset (2019)	Particle Swam Optimization	GPA & Attendance	Accuracy = 100%	(Sari & Sunyoto, 2019)
2020	University Dataset (2020)	-	Academic Performance	Accuracy = 79%	(Olalekan et al., 2020)
2020	Primary Schools Dataset (2017)	Connection Weights Algorithm	Gender, Parents' Qualification, Study Resources, Learning Habits & Academic Results	Accuracy = 84.5%	(Çevik & Tabaru-Örnek, 2020)

Table 3 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2020	OULAD (2015)	-	Demographics, LMS Clicks & Assessments' Scores	F1-Score = 0.96	(Tomasevic et al., 2020)
2019	University Dataset (2014–2018)	Pearson Correlation	GPA, Admission Scores, Admission & Current Study Years	Accuracy = 83%	(Mutanu & Machoka, 2019)
2019	Schools Dataset (2005–2006)	Binary Genetic Algorithm	Demographics, Family, Health, Social Activities, Attendance & Time Spend on Different Activities	Accuracy = 95%	(Turabieh, 2019)
2020	Chinese Public University Dataset (2020)	RankNet	Behavioral Features	Accuracy = 82%	(Zong et al., 2020)
2020	-	Information Gain	Personal Information, Domestic & School Factors	Accuracy = 99%	(Sokkhey & Okazaki, 2020a, 2020b, 2020c)
2019	College Dataset (2019)	-	Assessments' Marks	Accuracy = 95%	(Hussain et al., 2019)
2019	University Dataset (2015–2018)	-	Basic Information, Course, Program Year & GPA	Accuracy = 95%	(Amazona & Hernandez, 2019)
2019	Kaggle Dataset (2016)	-	Interaction with Learning Resource, Parents' Motivation, Demographics, Attendance & Grades	Accuracy = 84%	(Vijayalakshmi & Venkatachalapathy, 2019)
2020	High School in Kazakhstan Dataset (2012–2013)	-	Family Details, Teacher' Motivation & Grades	Accuracy = 91%	(Karlık & Karlık, 2020)
2020	College Dataset (2020)	-	Demographics, Study Habits & Social Interactions	MSE = 13.63	(Liu et al., 2020)
2018	High Schools Dataset (2018)	-	LMS Exercises	Accuracy = 76%	(Su et al., 2018)
2018	University Dataset (2007)	-	Students' Socioeconomic Background & Academic Performance	Accuracy = 95%	(Cavazos & Garza, 2017)
2019	University Dataset (2018)	-	Basic Information & GPA	Accuracy = 73%	(Umar, 2019)

Table 3 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2019	MOOC Dataset (2019)	-	Assignments	Accuracy = 70%	(Qu et al., 2019)
2020	University Dataset (2020)	-	Students' Interactions with Programming Software	Accuracy = 89%	(Pereira et al., 2020)
2020	University Dataset (2020)	-	Academic Performance	Accuracy = 70%	(Bravo et al., 2020)
2020	University Dataset (2018–2019)	-	Secondary Education, Subjects, Study Time & Computer Skills	RMSE = 2.42	(Khazaaleh, 2020)
2020	University Malaysia Pahang (2002–2015)	-	Gender, State & Entry Qualification	Accuracy = 92%	(Yahaya et al., 2020)
2018	University Dataset (2007–2008)	-	Exam Scores	Accuracy = 95%	(Iyanda, 2018)
2018	University of Gujrat (2007)	-	Demographics, Study Habits, Learning Interaction & Exam Performance	Accuracy = 95%	(Ahmad & Shahzadi, 2018)
2019	College Dataset (2016–2017)	-	Assessments, Number of Emails & Posts	Accuracy = 97%	(Altaf et al., 2019)
2019	University Dataset (2012–2017)	-	Number of Attempted Quizzes, Scores, Submission Time & Mid-Term Scores	Accuracy = 65%	(Sukhbaatar et al., 2019)
2019	Moodle Dataset (2019)	-	LMS Activities	Accuracy = 91%	(Raga & Raga, 2019)
2019	College Entrance Dataset (2019)	Spearman's Correlation	Gender, Age, Admission Scores & Learning Attitude	Relative Error = 0.1	(Liu, 2019)
2019	University Dataset (2016)	Spearman's Correlation	Grades & University Internet Usage	Accuracy = 70%	(Xu et al., 2019)
2019	University Dataset (2019)	-	Background Information, Attendance, Classroom Leadership & Admission, Assignments & Quizzes Scores	RRMSE = 0.4	(Mi, 2019)

Table 3 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2020	OULAD (2015)	-	Demographics, VLE Interactions & Academic Performance	Accuracy = 90%	(He et al., 2020)
2020	Questionnaire (2020)	-	Attendance, Study Hours & Marks	AUC = 0.86	(Makombe & Lall, 2020)
2020	OULAD (2015)	Sparse Feature Reduction	Demographic, VLE Clicks and Assessments' Scores	Accuracy = 93%	(Waheed et al., 2020)
2020	OULAD (2015)	-	Learning Behaviors	F-measure = 80%	(Karimi et al., 2020)
2020	Kaggle Dataset (2019)	Cluster-based Linear Discriminant Analysis (CLDA)	Demographic, Attendance& Academic Features	Accuracy = 93%	(Sood & Saini, 2020)
2020	OULAD (2015)	LSTM	Demographic & Learning Behavior	Accuracy = 61%	(Li, 2020)
2020	Saudi Public University (2016–2019)	-	Academic Records	Accuracy = 79.22%	(Mengash, 2020)

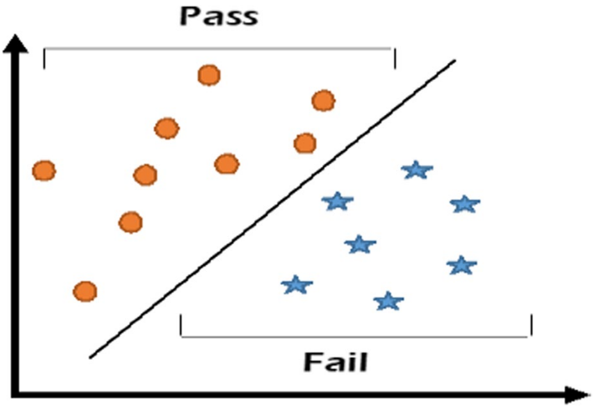

Fig. 4 Support Vector Machine (SVM)

2020). An ensemble model combines the results of different data mining techniques to make prediction more accurate. Such a hybrid approach is used (Kamal & Ahuja, 2019) by combining the prediction results of Decision Tree, Naïve Bayes and SVM. The study shows that ensemble model achieved an accuracy of 98.5%. A study (Wu et al., 2019b) proposed deep Neural Network prediction model based on CNN, Long Short-Term Memory (LSTM) and SVM models, and proved that hybrid model predicts more accurately (i.e. F-measure = 95.03%) as compared to linear SVM (i.e. F-measure = 92.48%). The prediction results of Deep Belief Network and SVM models are hybridized to decrease variance, and to enhance prediction results (Vora & Rajamani, 2019).

A comparative study is proposed to predict computer science graduation students' Grade Point Average (GPA), based on ANN, SVM and extreme learning machines as prediction models. The study concluded that students' GPA in previous semesters is the best indicator of their success or failure in final year exam, however, SVM model achieve highest prediction results (accuracy = 97.98%), followed by extreme learning machines (accuracy = 94.92%) (Tekin, 2014). Different studies have been proposed which compared SVM with other prediction models. Decision Tree (Naicker et al., 2020), Random Forest (Lottering et al., 2020), Logistic Regression (Aluko et al., 2016; Bhutto et al., 2020; Heuer & Breiter, 2018), Naïve Bayes (Soni et al., 2018; Fachrie, 2019), Random Forest, Neural Network (Solís et al., 2018; Ahmed et al., 2020b), KNN (Wiyono et al., 2020) and MLP (Zohair, 2019) are compared with SVM prediction model. In all the mentioned studies SVM achieved better accuracy as compared to other prediction models, applied on different students' attributes at different education levels.

Educational datasets consist of large databases with number of students' attributes and details. Not all the attributes influence their exam performance, therefore, all students' attributes cannot be used in prediction model. To select most influencing features, ensemble feature selection technique has been used. The study shows that SVM model with selected feature gave better accuracy as compared to prediction with random features (Lu & Yuan, 2018). In (Zaffar et al., 2020) correlation-based filtering is used to select most significant features for prediction process.

Features based SVM model achieved a F-measure of 90%. Principal Component Analysis (PCA) is used to explore correlation between students' social activities and their scores in English. The prediction model (i.e. SVM) shows that finding correlation between students' attributes increases prediction results (Zhao et al., 2020b). Open University Learning Analytics (OULA) dataset is one of the mostly used datasets in educational research. Datasets consist of students' demographic data, number of clicks, and assessments marks. This dataset is used to build SVM prediction model which forecast 93.5% accurate results (Chui et al., 2020). A prediction model may achieve different results when operated on different input features. Therefore, students' factors that influence their academic performance play a major role in prediction. A study examined students' MOOC dataset and found that students' performance in semester exercises is the best predictor followed by their clicks and interaction with learning material (Moreno-Marcos et al., 2020). Another study proved that using all students' data sources e.g., survey data, academics, interaction with learning resources doesn't provide most accurate results. It is suggested to combine only significant features for students' academic results prediction. The above three mentioned studies also proved that SVM is the prediction model for students' academic results (Yu et al., 2020). Table 4 presents a summary of research studies implementing SVM for students' performance prediction.

3.5 E. K-Nearest Neighbors

K-Nearest Neighbor (KNN) is a similarity approach. It stores data based on their similar attributes. This technique assumes that data items with similar attributes are most probably placed in the same class, where "K" is number of nearest neighbors that are selected to predict the class of an unknown object. When a new unknown data item is to be placed in a class, k nearest neighbors is selected based on shortest distance between new item and its neighbors. The new item is given the label of class which has majority of the nearest neighbors (Sen et al., 2020). Figure 5 presents a KNN prediction model with 3 nearest neighbors.

Five DM algorithms are compared namely Naïve Bayes, Decision Tree, KNN, C4.5 and SVM to generate best prediction model for students' exam performance prediction and found that KNN outperform other classification model with a highest accuracy of 100% (Vital et al., 2021). A study is conducted to compare four different data mining algorithms for students' academic performance and found that KNN gave best results as compared to other prediction models (Kulkarni & Ade, 2014). Students' learning behaviors are used in KNN based prediction model. It is found that students' clickstream data is very useful to predict their results (Brinton & Chiang, 2015). KNN algorithm with fixed and random number of 'k' is applied with ensemble clustering techniques. Students' demographics, enrollment and performance records are used to prediction of final exam outcome (Iam-On & Boongoen, 2017). A fast KNN algorithm is proposed to decrease model's processing time without compromising prediction accuracy. The proposed model gave better accuracy i.e., 96.6% as compared to traditional KNN model. The proposed model also decreases processing time up to 90% (Ahmed et al., 2020c). KNN is used to predict

Table 4 SVM for students' performance prediction

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2013	University Database (2013)	-	GPA & Subject Scores	Accuracy = 91.1%	(Huang & Fang, 2013)
2014	University Students' Data, USA (2014)	-	Subjects' Grades	Accuracy = 50%	(Mativo & Huang, 2014)
2014	Student Affairs of Fırat University, Turkey (2006–2010)	-	Scores	Accuracy = 97.98%	(Tekin, 2014)
2016	University of Lagos, Nigeria Dataset (2013–2014)	-	Demographics, Basic Information, Family Details & Academic Performance	Accuracy = 80%	(Asogbon et al., 2016)
2017	University Database (2017)	Information Gain	Personal, Admission Data, Exercise Details & Scores	F-Measure = 0.9	(Costa et al., 2017)
2018	School Dataset (2018)	Ensemble Approach (Pearson Correlation Coefficient & Euclidean Distance)	Demographics, Family Details, Social Activities, Health, Term Grades & Attendance	Mean Accuracy = 93.5%	(Lu & Yuan, 2018)
2018	OULAD (2015)	-	Demographics, Academic Performance & Number of Clicks on LMS	F-Measure = 91%	(Heuer & Breiter, 2018)
2018	OULAD (2015)	-	Demographics, Sum of Clicks & Assessments Scores	Accuracy = 94%	(Chui et al., 2020)
2018	Olabisi Onabanjo University Dataset (2018)	-	CGPA	Accuracy = 76%	(Aluko et al., 2016)
2018	University Dataset (2017–2018)	Decision Tree, SVM & Naïve Bayes	Attendance, Marks, Family Details, Interest in Studies, Extra-Curricular Activities & Habits	Accuracy = 83.33%	(Soni et al., 2018)
2018	University Dataset (2011–2016)	-	Socio-Demographics, Year of Enrollment, Enrolled Program & Academic Record	Specificity = 94%	(Solís et al., 2018)
2019	University Dataset (2019)	Hierarchical Clustering Method	Age, Teacher Name, bachelor's degree Name, GPA & Selected Courses	Accuracy = 76%	(Zohair, 2019)

Table 4 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2019	College Dataset (2019)	Information Gain & Entropy	Demographics, Academic Performance & Social Activities	-	(Kamal & Ahuja, 2019)
2019	KDD Cup (2015)	CNN & LSTM	Learning Activity Records	F1-Score = 95.03%	(Wu et al., 2019b)
2019	Questionnaire Based Dataset (2019)	-	Psychological Factors	Accuracy = 90%	(Burman & Som, 2019)
2019	-	PCA	Basic Information, Academic Details & CGPA	Accuracy = 71%	(Vora & Rajamani, 2019)
2020	High School Dataset (2019)	-	Personal Information, Lunch Access, Parents' Education & Academic Grades	Accuracy = 90%	(Naicker et al., 2020)
2020	Kalboard 360 Dataset (2015)	Gain Ratio	Demographics, Academic Factors, Parents Involvement & Students' Interaction With LMS	Accuracy = 79%	(Bhutto et al., 2020)
2020	Three Datasets of Schools & Colleges (2020)	Fast Correlation-Based Filter	Demographics, Academic & Behavioral Features	F-Measure = 90%	(Zaffar et al., 2020)
2020	MOOC Dataset (2016)	-	LMS Activities & Assessments	MAE = 0.07	(Moreno-Marcos et al., 2020)
2020	University Dataset (2006–2016)	Data-Based Sensitivity Analysis	Demographics & Academic Details	AUC = 0.93	(Gil et al., 2020)
2020	University Dataset (2013–2017)	PCA	Demographics, Financial Aid & Academic Performance	Accuracy = 99%	(Lottering et al., 2020)
2020	STEM Students' Dataset (2016–2018)	-	Institutional Factors, LMS Activities & Students' Learning Behavior	Accuracy = 67%	(Yu et al., 2020)
2020	-	-	Residence, School, Parents' Job & GPA	Accuracy = 95%	(Wiyono et al., 2020)
2020	-	PCA	Students' Internet Details	F1-Score = 78%	(Zhao et al., 2020b)
2020	University Dataset (2020)	Gain Ratio, Correlation & CFS Subset	Personal Factors, Academic Performance & Internet Usage	Accuracy = 99%	(Ahmed et al., 2020b)
2020	University Dataset (2014)	-	Academic Details	Accuracy = 92%	(Fachrie, 2019)

different courses' results and gained 73.33% accurate results (Aluko et al., 2016). Another study proves that feature selection methods improve KNN prediction results (Ahmed et al., 2020a). Table 5 presents a summary of research studies implementing KNN for students' performance prediction.

3.6 F. Random Forest

Random Forest (RF) is an ensemble machine learning algorithm, comprised of multiple Decision Trees. Each Decision Tree is built using different set of training features, and at last, the prediction results of all Decision Trees are merged to achieve a more accurate prediction result. The final class with majority votes is selected as the predicted class (Bruce, 2019). Figure 6 describes the working of Random Forest classification model.

A research study focuses on the development of a Random Forest based prediction model for students' learning outcomes. Students' interactions with e-learning management system e.g. students' visit, resource views, assignments submission and scores are used to identify key attributes with an aim to achieve highest prediction accuracy. Random Forest gave 76.9% accurate results (Abubakar & Ahmad, 2017). A two-step prediction process is performed, at first step classification algorithms are compared to predict students' dropout and secondly students' grades are predicted using regression analysis. For classification, Random Forest gave highest accuracy i.e. 82% (Rovira et al., 2017). Another two-step model is proposed for students' prediction. At first, Random Forest is used to assign weights to each attribute based on their contribution to final grades prediction. Most correlated attributes are then used in prediction model and 96.1% is achieved (Miguéis et al., 2018). Fourteen data mining algorithms are compared to find best prediction model based on students' demographic and academic attributes. The study found that Random Forest gave highest accuracy i.e. 93% (Senthil & Lin, 2017). An automatic prediction model was built using clustering and classification models for predicting students' promotion in the next class. Dataset consisting of 151 attributes was collected from different universities and colleges. Most relevant attributes were selected using k-means clustering and an ensemble voting based techniques was used to predict students' outcome in final examination. Results shows 87.5% accuracy when classify the raw data, however, 96.78% was achieved after most relevant attributes selection (Thakar & Mehta, 2017). Random Forest is used to find most influencing factors for students' exam performance prediction. The proposed study found that students' previous semesters' CGPA and interaction with learning resources are best predictors of students' final results (Sandoval et al., 2018). Another research study combined Relief-F and Random Forest models for selecting most significant attributes for students' final exam scores prediction. The proposed model gave 97.88% prediction accuracy and shows that students' attendance, extra-curricular activities, previous grades and parents' education are highly influencing students' exam results, therefore can be used for exam performance prediction (Deepika & Sathyanarayana, 2019). A research study performed a comparison of four data mining algorithms namely Decision Tree, PART, Random Forest & Bayesian Network. Gain Ration, Information Gain

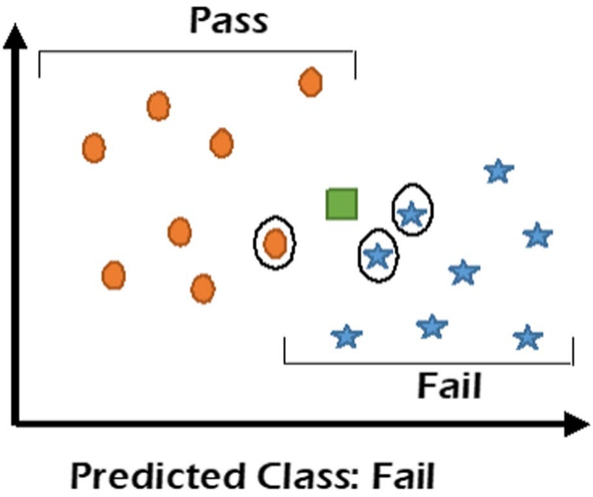

Fig. 5 K-Nearest Neighbors (KNN)

& Relief-F are used to select most significant features. Random Forest with twelve students' attributes gave best results i.e. 99% accurate results (Hussain et al., 2018). Another study used Random Forest to classify students into dropout and non-dropout students and found an F1-score of 60% (Polyzou & Karypis, 2018).

Above studies proposed prediction model based on number of attributes, however some of the e-learning institutes have very limited attributes that can be used for prediction. Random Forest is used for prediction performed using lectures views, resources' access and assessments' scores. The proposed model gave 84% accuracy and proved that students' LMS interactions and grades can be used for performance prediction (Wakelam et al., 2020). Three data mining algorithms are compared for modeling students into pass/fail segments. Random Forest outperform with 95.45% as compared to KNN and Naïve Bayes (Lenin & Chandrasekaran, 2019). Similarly, another study compared six data mining techniques and found that Random Forest gave highest results i.e. 94.1% (Rifat et al., 2019). To early predict students' results, features that are highly correlated with results are extracted and results are predicted using Random Forest. The studies show that not all attributes contribute in prediction process, however, using irrelevant attribute may negatively affect prediction results (Masood et al., 2019; Nuankaew & Thongkam, 2020). Different machine learning algorithms are used for students' final grades prediction. The study found that Random Forest gave best results as compared to other algorithms, when applied with students' CGPA, attendance and extra-curricular activities (Singhani et al., 2019). As early students' results are predicted, the more chances students will have to prevent their dropout. A Random Forest based phased prediction model is developed to predict students' dropout at different stages of their semester. The proposed study shows that students' results can be predicted using demographic data at the start of semester, while in the mid of semester predicted results will become more accurate by adding students' learning behaviors (Chen et al., 2020).

To check the influence of different features on students' dropout prediction, Random Forest model is used to check correlation of all features with results. The

Table 5 KNN For Students' Performance Prediction

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2014	University Database (2014)	Non-Nested Generalization (NNGE)	Academic Factors E.G. Marks, Attendance	Accuracy = 99%	(Kulkarni & Ade, 2014)
2015	MOOC (2015)	-	Video Clickstream Data	Accuracy = 75.9%	(Brinton & Chiang, 2015)
2015	Mae Fah Luang Thailand University, Dataset (2009–2012)	Weighted Connected Triple Algorithm	Gender, Province, Entry Type & GPA	Error Rate = 0.073	(Iam-On & Boongoen, 2017)
2020	UCI Repository Dataset (2014)	-	Demographics, Health, Social & Academic Performance	Accuracy = 96%	(Ahmed et al., 2020c)
2020	Northwestern University, Bangladesh (2020)	Genetic Algorithm	Attendance & Academic Scores	Accuracy = 91.37%	(Ahmed et al., 2020a)
2020	Olabisi Onabanjo University (2018)	-	Exam Grades	Accuracy = 73.33%	(Aluko et al., 2016)

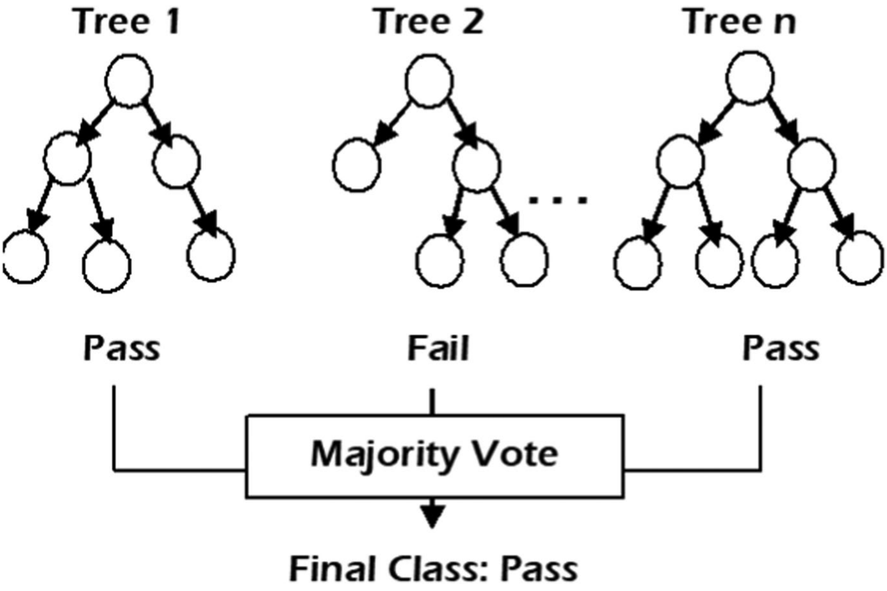

Fig. 6 Random Forest (RF)

proposed study shows that students' gender, age and region are negatively correlated with prediction model and their removal increased the accuracy of the model (Bruce, 2019). Four data mining prediction techniques namely Naïve Bayes, Logistic Regression, KNN and Random Forest are compared, and the study found that Random Forest gave best results. In the proposed study, dataset was divided into 10 weeks in a way that attributes are added to the dataset according to their availability. A comparative study is conducted to find best prediction model among Statistical techniques, Deep Belief Learning and Machine Learning algorithms. The study found that overall Machine Learning techniques gave best results and among ML techniques, Random Forest outperform with an accuracy of 93.5% (Sokkhey & Okazaki, 2019). A research study is conducted to predict students' results without knowing their previous grades. The proposed study shows that students' grades have great influence on results' prediction, however their demographic attributes can give enough accuracy for preventing students' dropout (Rajak et al., 2020). Ensemble based techniques are used for prediction of at-risk students, and found that Random Forest outperform other techniques (Kaviyarasi & Balasubramanian, 2020). Random Forest was applied on a dataset gathered from Open Universities of China including students' demographics, behavioral and academic performance data was used. Before prediction, regression analysis was performed to find correlation between students' attributes and final exam grades. Results shows that students' views, clicks and learning activity duration are best predictor for their future grades (Narayanasamy & Elçi, 2020). Different data mining algorithms Naïve Bayes, ANN (Adekitan & Salau, 2020), Logistic Regression (Alhassan, 2020), Decision Tree (Farissi & Dahlan, 2020), SVM (Sokkhey & Okazaki, 2020a, 2020b, 2020c), KNN

(Rincón-Flores et al., 2020; Sokkhey & Okazaki, 2020a, 2020b, 2020c) are compared with Random Forest, and it is found that Random Forest gave best prediction results. Table 6 presents a summary of research studies implementing Random Forest for students' performance prediction.

4 Students' attributes used for academic performance prediction

Same data mining techniques gave different prediction results when used in different research papers as shown in the above tables. This difference in results is observed because each author used a different set of students' attributes as input. The major challenge in developing a prediction classifier is the input data type. Some students' attributes may have more impact on prediction results as compared to the other attributes. Some datasets used in literature studies are based on distance learning while the others are traditional classroom datasets. The use of different students' attributes makes same algorithm gave a different result in the terms of accuracy. Some of the commonly used attributes are students' grades achieved in a previous exam, attendance of the same course, gender, age, place of residence, family, social activities, learning behaviors, and interaction with learning resources etc. An analysis of research studies shows that students marks are the most used as it shows academic potential of a student. A study on students' progress pattern shows that students who outperform in a midterm exam are most likely to show good results in their final exam too. Similarly, students who gain less marks in the start of a degree, do not show any progress in their results till the end of degree program. Most of the students tend to remain in the same category (Asif et al., 2017). Students' attendance is second mostly used attribute which gave students' performance prediction. Students who attend more lectures have more chances to pass their examination (Hughes & Dobbins, 2015). Students' demographic attributes highly affects their academic performance.

In a research study, state-of-the-art regression algorithms are applied to predict students' exam performance. A total of 354 graduate and post graduate students' records are collected from Hellenic Open University database. Along with students' previous academic performance, demographic records including age, gender, marital status, jobs and number of children are considered for prediction process. Statistical measure i.e., Relief F is used to find most influencing features in dataset. After feature ranking, M5rules gave minimum error rate (Kotsiantis & Pintelas, 2005). A students' performance model is proposed with tenfold cross validation. Four Bayesian algorithms namely Naïve Bayes, AODEsr, WAODE and HNB are compared. The study concluded that AODEsr outperformed i.e. 64.6% accurate results when applied with students' academic performance and co-curricular activities (Sundar, 2013). Matrix Factorization algorithm applied in (Sweeney et al., 2015), gave best results for students' next term grades prediction i.e. RMS = 0.775. Three classification algorithms are compared to find best prediction model of final exam grades. The results proved that Rule based model is the best prediction model and students' demographics and learning behaviors can be used to generate prediction results (Ahmad et al., 2015). Another research study applied deep neural

Table 6 Random forest for students' performance prediction

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2017	University of Barcelona Dataset (2009–2014)	-	Exam Grades	Accuracy = 91%	(Rovira et al., 2017)
2017	UCI Repository Dataset (2014)	Leave-One-Out Method	Demographics, Family Details, Activities, Previous Performance & Attendance	Accuracy = 93.3%	(Senthil & Lin, 2017)
2017	Moodle LMS (2015)	-	Courses, Assignments, Learning Resources & Exam Scores	Accuracy = 92.3%	(Abubakar & Ahmad, 2017)
2017	University Dataset (2017)	Chi-Square	Demographics, Academic Performance, Personality & Social Details	Accuracy = 96.78%	(Thakar & Mehta, 2017)
2018	University LMS: SAKAI & DARA (2013–2014)	-	Demographics, Academic Information & LMS Activity Logs	RMSE = 0.08	(Sandoval et al., 2018)
2018	UCI Repository Dataset (2014)	Relief-F & Budget Tree	Personal, Family, Social Activities, Health, Extra-Curricular Activities, Previous Failures & Attendance	Accuracy = 97.8%	(Deepika & Sathyanarayana, 2019)
2018	University Dataset (2003–2015)	Random Forest	Socio-Demographic Information, Entry Scores & High School Grades	Accuracy = 96%	(Miguéis et al., 2018)
2018	College Dataset (2018)	Correlation-Based Attribute Selection	Demographics, Family Information, Study Time, Attendance & Academic Percentage	Accuracy = 99%	(Hussain et al., 2018)
2018	University of Minnesota Dataset (2018)	-	Students' Basic Information & Courses Details	F1-Score = 0.61	(Polyzou & Karypis, 2018)
2019	-	-	VLE Access, Attendance & Assessments' Scores	Accuracy = 98%	(Wakelam et al., 2020)
2019	Martin Luther Christian University Dataset (2019)	Random Forest	Demographics, Courses & Scores	Accuracy = 95%	(Lenin & Chandrasekaran, 2019)

Table 6 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2019	University in Bangladesh Dataset (2013–2016)	-	GPA	Accuracy = 94%	(Rifat et al., 2019)
2019	Two Kaggle Datasets (2017)	-	Demographics, Social Activities, Learning Resources, Health, Family Details, Attendance & Assessments	Accuracy = 99%	(Masood et al., 2019)
2019	University Dataset (2019)	-	Attendance, Grades, CGPA, Extra-Curricular Activities & Feedback	Accuracy = 57%	(Singhani et al., 2019)
2019	XJTUDLC Dataset (2015)	Chi-Square Correlation	Demographics, Semester & LMS Behaviors	Precision = 93.6%	(Chen et al., 2020)
2019	OULAD (2015)	Random Forest	Demographics, Academic Scores & LMS Clicks	Accuracy = 93%	(Bruce, 2019)
2019	Xorro-Q Educational Tool (2019)	-	Students' Academic Activities & Scores	F-Measure = 0.88	(Ramaswami et al., 2019)
2019	-	Entropy & Information Gain	Domestic, Academic & Personality Factors	Accuracy = 80%	(Sokkhey & Okazaki, 2019)
2020	UCI Repository Dataset (2014)	-	Demographics, Health, Social Activities, Attendance & Grades	Accuracy = 99%	(Rajak et al., 2020)
2020	-	-	Mobile Access, Assessments, Alcohol Consumption, Family Size & Extracurricular Activities	Accuracy = 85%	(Kaviyarasi & Balasubramanian, 2020)
2020	MOOC (2015)	Clustering	Student Characteristics, Activity Logs & Academic Performance	Accuracy = 96.39%	(Narayanasamy & Elçi, 2020)
2020	Rajabhat Maha Sarakham University, Thailand (2020)	Gain Ratio	Demographic Attributes & GPA	F-measure = 94.7%	(Nuankaew & Thongkam, 2020)

Table 6 (continued)

Year	Dataset Used	Feature Selection	Students' Attributes	Highest Results	Ref
2020	Nigerian university (2020)	-	Ethnicity	F1-score = 79%	(Adekitan & Salau, 2020)
2020	King Abdulaziz University (2017–2019)	Filter and Wrapper	LMS Activities & Assessments' Marks	F-measure = 99%	(Alhassan et al., 2020)
2020	Kaggle Repository (2016)	Genetic Algorithm	Demographic, Behavioral & Academic Features	F-measure = 81.18%	(Farissi & Dahlan, 2020)
2020	University Dataset (2020)	-	Students & Teacher Interviews	-	(Rincón-Flores et al., 2020)
2020	Schools Dataset (2020)	PCA	Family, School & Personality Factors	Accuracy = 99.72%	(Sokkhey & Okazaki, 2020a, 2020b, 2020c)
2020	High Schools Dataset, Cambodia (2020)	Mutual Information & Chi-Square	Family, School & Personality Factors	Accuracy = 99.98%	(Sokkhey & Okazaki, 2020a, 2020b, 2020c)

network to predict students' final exam grades. MOOC dataset is used for building prediction model, consisting of students' interactions with learning materials and activity logs. Deep learning shows best results when compared to baseline algorithms (Wang et al., 2017). Five ML algorithms namely generalized linear model, MLP, Random Forest, gradient boosting tree and ANN are compared to find best prediction model for students' exam scores. Dataset consisting of students' assessments scores was collected from DIT University, Dehradun. A highest accuracy of 98.26% is achieved by gradient boosting model (Kumar & Garg, 2019). A hybrid of seven classification algorithms namely SVM, KNN, Decision Tree, AdaBoost, MLP, Extra Tree and Logistic Regression is used to predict students' scores using their institutional dataset attributes. The proposed weighted voting approach shows better results i.e. 81.37%, as compared to individual algorithms (Zulfiker et al., 2020).

An ensemble algorithm is proposed comprised of WINNOW, 1NN and Naïve Bayes. The proposed ensemble model receives input features and predict outcome based on a majority vote. Hellenic Open University dataset consisted of students' assignments scores are used to classify students into pass and fail. Proposed ensemble model gave best accuracy i.e. 78.95% as compared to individual supervised learning models (Kotsiantis et al., 2010). Another ensemble algorithm is proposed based on three state-of-the-art classifiers namely J48, IBK and AODE. A majority vote is received by implementing three algorithms in a single model and it is found that ensemble approach gave 85% accurate results. The study used a combined dataset of academic and demographic attributes (Pandey & Taruna, 2018). Ensemble approach based on Stacked generalization is used to predict students' performance based on demographics, psychological, personality and institutional attributes (Adejo & Connolly, 2018). The proposed study found that a hybrid of Decision Tree, ANN and SVM gave better accuracy as compared to individual results. Features optimization using genetic algorithm is applied with supervised learning algorithms gave 75.55% accuracy (Pereira et al., 2019). In (Mi et al., 2018) Genetic Algorithm is used to develop an early warning system for students. The main significance of proposed model is that it gives a clear description of prediction process by using if–then rules. Similarly, another research study used genetic programming for students' success prediction in online courses. Pearson's correlation is used to find attributes that highly contribute to final grade prediction. It is found that students' scores are most significant features with a correlation coefficient r = 0.78 (Ulloa-Cazarez et al., 2018). These studies prove that feature selection algorithms enhance prediction accuracy and reduce model computational time.

In (Abu Tair & El-Halees, 2012), association rules, classification and clustering algorithms are applied on students' demographic and academic datasets, to predict students' dropout in college degree. An accuracy of 78.95% is gained by proposed association and classification rules. The proposed model presents that students' gender, specialty and scores in secondary school are highly correlated with their semester results. Another study proposed a Neuro Fuzzy based classification model. The proposed model with threefold cross validation gained best results i.e., RMSE = 0.256. The study shows that students intelligence, motivation and interests in studies can be used to predict their final exam performance (Hidayah et al., 2013). In (Márquez-Vera et al., 2016), Classification Rule Mining algorithm is proposed to predict students' who are more prone to dropout. If–then rules give a detailed vision of attributes that leads to final prediction. The proposed model shows that students final grades can be predicted in first 6 weeks of registration. If–then rules are generated for early prediction of students' results. To enhance prediction accuracy, rough set theory is used for data dimensionality reduction which shows 79.23% accuracy (Sudha & Kumaravel, 2017). In (Czibula et al., 2019), Relational Association Rules is used for predicting students' grades in final semester. Data was collected from Babes-Bolyai University consisting of students first three semester GPA. The proposed model gave best results i.e., F-measure = 0.84. Apriori algorithm is used (Anwar & Rani, 2020) to classify students into dropout and no dropout classes for their future results. Students' previous exam scores in Mathematics are used as predictors of future Mathematics scores. Findings revealed that students with higher scores in prerequisite classes are more likely to have better performance in next classes. M5 Rules Algorithm is implemented (Chand et al., 2020) for grades prediction. Using scores of different subjects, M5rules gained highest accuracy i.e., 89.2% as compared to Random Forest and Linear Regression.

Different students' attributes are available at different phases. Two datasets are used consisted of different attributes, first with demographics and previous class performance factors that are available at the beginning of session, while second dataset includes demographics as well as students' assessments scores, attendance and subjects. Research concluded that students' neighborhood, age, assessments scores and attendance are highly correlated with final exam grades (Fernandes et al., 2019). Students' behavioral attributes e.g. orderliness inside the institute is found as most significant attribute to predict their academic performance (Cao et al., 2018). In another study, students' social media activities are examined to find impact of students' academic and non-academic social media activities on their final exam scores. Findings revealed that students' social media activities can be used for predicting their final exam performance (Chang et al., 2019). Learning strategies and motivation are found as most significant attributes for students' CGPA prediction, with a correlation of 0.243 and 0.193 respectively (Nabizadeh et al., 2019). Students' response time is found as a good predictor of students' scores, as minimum response time shows students' knowledge and attention towards the lecture. Additive Factors Analysis approach predict students' results with 87.8% accuracy (Chounta & Carvalho, 2019). In online courses, students' interactions with learning resources are found as a significant variable for predicting their academic performance. More than 2000 websites frequently visited by students, are considered for research. Research results shows that websites containing videos, games and music are negatively correlated, however, visiting learning based websites are positively correlated with academic performance (Wu et al., 2019a). Similarly, in another study, students' interactions are considered to predict their final exam scores. OULAD dataset provides details of learning resources and sum of clicks performed by students during their course. Long short-term memory algorithm shows 59% precision in the 1st week while 93% precision was achieved in the last week of course (Aljohani et al., 2019). Canonical Correlation Analysis is used to explore relation between different learning resources. The proposed study found that students' performance in one learning resource can be used to predict their performance in other type of learning resources (Sahebi & Brusilovsky, 2018). Another research study aimed to explore potential attributes for students' performance prediction. Out of 45 students' attributes, research study found that students' previous grades, attention in class, study-room and extra-curricular activities have positive while access to mobile phone, alcohol consumption and more travel time to school have negative impacts on students' academic performance (Kaviyarasi & Balasubramanian, 2018). Logit leaf model is implemented for students' performance prediction in online courses. Over 10,554 students' records were used comprised of their learning patterns and activities. The study revealed that students' academic engagement is the best predictor of students' academic performance (Coussement et al., 2020). Students' final grades are predicted using 2500 students' data registered in different courses. Rule Induction classifier gave 96.25% accuracy (Majeed & Junejo, 2016). Input–Output Hidden Markov Model is proposed to predict students' performance using students' weekly activities in online learning environment. The proposed model gave 82% accuracy in the second week, while 84% accuracy is gained in the last week of course (Mubarak et al., 2020). Above studies shows that students' academic records, demographics and learning behavior are the best predictor. Several studies proved that using data preprocessing and feature selection techniques enhances the prediction results.

Different datasets are used in research studies, most of researchers gather data from schools, colleges and universities' databases, LMS systems or conducted surveys to collect students' responses. Several studies used online available datasets. Four publicly available datasets are mostly found in the research papers for students' performance prediction named as: OULAD, MOOCs, Moodle and UCI Repository dataset. Open University Learning Analytics Dataset[1] (OULAD) contains a data of 22 courses and 32,593 students. Students' demographic attributes, sum of clicks and assessments' results are available with their final exam result. Massive Open Online Courses[2] (MOOCs) offer opportunities for distance learning. Modular Object-Oriented Dynamic Learning Environment[3] (Moodle) is another learning management system used for online learning courses. Students' LMS data is available and used in lot of studies to predict students' performance and their learning behaviors. UCI Machine Learning Repository offers a dataset[4] for students' performance prediction. The dataset contains 23 attributes including demographic, social and academic records. A dataset of 649 students from two secondary schools is available and used in different studies for final exam grade prediction. Table 7 presents a summary of different students' attributes for their exam performance prediction.

[1] https://analyse.kmi.open.ac.uk/open_dataset

[2] https://www.mooc.org/

[3] https://moodle.org/

[4] https://archive.ics.uci.edu/ml/datasets/student+performance

5 Tools used for data mining

This section presents a comparison of data mining tools used for students' academic performance prediction. There are several studies presented above, these research studies used different data mining tools for the prediction process. A wide number of tools are available to build prediction models using machine learning. These tools make it very easy to perform prediction tasks, data analysis, feature selections, data cleansing and building classification and regression models etc. This section presents data mining tools that are used in literature for the prediction of students' academic achievement i.e., RapidMiner, WEKA, MATLAB, and Python.

5.1 A. RapidMiner

RapidMiner provides a user-friendly interface to build prediction models. RapidMiner is very easy to use as it provides a graphical, code-free environment. Prediction models are built by drag and drop operations. All classification and regression models are available. To build a model, import data into RapidMiner Studio, set parameters and drag & drop required model into design screen. Resultant model will appear in results section. It also supports statistical analysis of results to evaluate the accuracy of model, visualization is also available to provide graphical representation of results. RapidMiner also provide step by step tutorials that are helpful for beginners (Osmanbegovic & Suljic, 2012).

5.2 B. WEKA

WEKA stands for Waikato Environment for Knowledge Analysis. WEKA is another data mining platform for building prediction models. WEKA provides graphical user interface as well as command line interface to implement data mining algorithms. WEKA allows users to use its provided operators or to implement their own java codes. It is used to solve all classification, clustering, feature selection, data processing and regression problems. It is an open source and freely available software which increases its number of users (Shahiri & Husain, 2015).

5.3 C. MATLAB

MATLAB stands for "MATrix LABoratory", developed and sold by Mathworks, Inc. MATLAB is also used for data science problem solving. It allows implementation of data mining algorithms for classification and prediction problems. It reduces data preprocessing time, filters noisy data, plot data into graphs to allow users visualize data patterns and build data mining models. It also provides analysis features to evaluate model results (Tomasevic et al., 2020).

Table 7 Students’ attributes used for exam performance prediction

Year	Dataset Used	Prediction Algorithms	Feature Selection	Students’ Attributes	Highest Results	Ref
2005	Hellenic Open University Dataset (2000–2001)	M5rules	R Relief	Demographics & Academic Performance	MAE = 1.2	(Kotsiantis & Pintelas, 2005)
2010	Hellenic Open University Dataset (2010)	Ensemble (Naïve Bayes, 1-NN & WINNOW Algorithms)	-	Assignments’ Marks	Accuracy = 78.95%	(Kotsiantis et al., 2010)
2012	College of Science and Technology –Khanyounis (1993–2007)	Association Rules, Classification & Clustering	Local Outlier Factor	Basic Information, Enrollment Year, Academic Performance & Location	Accuracy = 67.50%	(Abu Tair & El-Halees, 2012)
2013	University Database (2013)	Adaptive Neuro Fuzzy Inference System	-	Intelligence, Educational Interest, Motivation	RMSE = 0.256	(Hidayah et al., 2013)
2013	College Students’ Data (2012–2013)	Aodesr	-	Assignment & Mid-Term Scores	Accuracy = 64.6%	(Sundar, 2013)
2015	George Mason University Dataset (2014)	Factorization Machine Model	-	Academic results	RMSE = 0.775	(Sweeney et al., 2015)
2015	School Database (2006–2013)	Rule Based Classification Algorithm	-	Demographics, Family Background & Academic Record	Accuracy = 71.3%	(Ahmad et al., 2015)
2018	College Dataset (2018)	Ensemble Voting (Naïve Bayes, IBK & J48)	Chi-Square Based Ranker	Academic & Demographic Factors	Accuracy = 85%	(Pandey & Taruna, 2018)
2016	University Dataset (2018)	Interpretable Classification Rule Mining	Chi-Square, Information Gain, Gain Ratio and Relief F	GPA, Classroom Factors, Number of Friends, Social Activities & Scores in Different Courses	Accuracy = 99%	(Márquez-Vera et al., 2016)
2016	College Database (2009–2013)	Rule Induction Method	-	Courses, Quiz Type, Class Participation, Lab Task & Instructor Name	Accuracy = 96%	(Majeed & Junejo, 2016)
2017	KDD Cup (2015)	Conrec Network	-	Learning Activities	F1-Score = 92.41%	(Wang et al., 2017)

Table 7 (continued)

Year	Dataset Used	Prediction Algorithms	Feature Selection	Students' Attributes	Highest Results	Ref
2017	UCI Repository Dataset (2014)	Rough Set Theory & Decision Rules	Correlation Based Feature Selection	Demographics, Social Activities, Interest in Studies & Attendance	Accuracy = 79.23%	(Sudha & Kumaravel, 2017)
2018	UCI Repository Dataset (2014)	-	Information Gain	Parents' Education, Job & Financial Status, Syllabus, Internet, Health, Extra-Curricular Activities, Study Hours & Attendance	-	(Kaviyarasi & Balasubramanian, 2018)
2018	MOOC (2015)	Canonical Correlation Analysis	PCA	Students' LMS Discussions, Assignments & Quizzes	RMSE = 0.1	(Sahebi & Brusilovsky, 2018)
2018	University of The West of Scotland Dataset (2018)	Ensemble (Decision Tree, ANN & SVM)	PCA	Demographics, Psychological, Personality, Employment & Institute Details	RMSE = 0.396	(Adejo & Connolly, 2018)
2018	University Dataset (2018)	Genetic Algorithm	-	Attendance, Assignments Score, Gender, Admission Scores, Class Leadership & Courses	Accuracy = 72%	(Mi et al., 2018)
2018	University LMS Database (2018)	Genetic Programming	-	Forum Views, Messages, Resources View	R = 0.78	(Ulloa-Cazarez et al., 2018)
2018	Schools Dataset (2015–2016)	Gradient Boosting Machine	-	Students' Basic Information, School Details, Absences & Grades	-	(Fernandes et al., 2019)
2018	University Dataset (2018)	RankNet	Spearman Rank Correlation Coefficient	Students' Behavioral Factors I.E. Diligence & Orderliness	AUC = 0.685	(Cao et al., 2018)

Table 7 (continued)

Year	Dataset Used	Prediction Algorithms	Feature Selection	Students' Attributes	Highest Results	Ref
2019	Turkish University Dataset (2019)	-	-	Social Media Activities, Internet Usage & Exam Grades	RMSE = 0.073	(Chang et al., 2019)
2019	Shahid Beheshti University, Questionnaire Based Dataset (2019)	-	Correlation	Demographics, Department, Motivation, Learning Strategies & CGPA	R = 0.193	(Nabizadeh et al., 2019)
2019	Datashop Repository (2019)	Additive Factors Analysis Model	-	Assessments	Accuracy = 87.8%	(Chounta & Carvalho, 2019)
2019	University Dataset (2019)	Association Rule Mining	-	Academic Grades	F1-Score = 63%	(Czibula et al., 2019)
2019	Students' Programming Course Data (2019)	Genetic Algorithm	-	Student Programming Behavior	Accuracy = 77%	(Pereira et al., 2019)
2019	DIT University Dataset (2015)	Gradient Boost Model	-	Pre-Admission, Mid-Term & Assessments' Marks	Accuracy = 98%	(Kumar & Garg, 2019)
2019	University Weblogs (2018)	Long-Short Tem Memory	Random Forest	Visited Websites	F1-Score = 86%	(Wu et al., 2019a)
2019	OULAD (2015)	LSTM Deep Learning	-	VLE Interaction	Precision = 93%	(Aljohani et al., 2019)
2019	University Dataset (2019)	M5 Rules	-	Marks	Accuracy = 89%	(Chand et al., 2020)
2020	University Dataset (2020)	Apriori Algorithm	-	Mathematics Marks	Confidence = 100%	(Anwar & Rani, 2020)
2020	University in Bangladesh Dataset (2018–2019)	Ensemble Voting	-	Attendance, Reappearing, Quizzes, Mid-Term Grades & Presentation	Accuracy = 81.73%	(Zulfiker et al., 2020)
2020	OULAD (2015)	Input–Output Hidden Markov Model	-	Demographics, Clicks and Assessments' Scores	Accuracy = 84%	(Mubarak, 2020)
2020	-	Logit Leaf Model	-	Students' Learning Activities	SD = 0.845	(Coussement et al., 2020)

5.4 D. Python

Python is a programming language used for implementing machine learning algorithms. It is an open-source program, freely available for commercial uses. Different libraries are available in python to implement codes, e.g., Pandas for data preparation, Scikit-learn for machine learning, Plotly for data visualization, and Theano for mathematical expressions (Stančin & Jović, 2019).

Some research studies used other data mining tools, e.g., SPSS (Moseley & Mead, 2008), KNIME (Adebayo & Chaubey, 2019; Rifat et al., 2019), R Studio (Kumar et al., 2019; Sukhbaatar et al., 2019; Lottering et al., 2020; Olalekan et al., 2020) and R Programming (Akçapınar et al., 2019; Figueroa-Cañas & Sancho-Vinuesa, 2020; Lenin & Chandrasekaran, 2019; Vijayalakshmi & Venkatachalapathy, 2019). Table 8 presents a summary of four frequent data mining techniques used for students' performance prediction.

6 Results and discussion

This section presents an overview of the research findings. Figure 7 clearly describes that students' performance prediction is of high interest in the present decade. Educational data mining is a new research domain but is rapidly growing because of its impacts and benefits gained by institutions. Figure 7 shows that Decision Tree is mostly used since last ten years but ANN, SVM and Random Forest are trending algorithms in the past three years. Below figures clearly present that the work on students' exam performance prediction is growing rapidly year by year. However, different studies used different techniques to improve the prediction results. Figure 8 gives an overview of frequently used data mining techniques for the prediction of students' final exam performance in the last years. The mostly used techniques are Decision Tree and ANN. While least used technique is KNN. Decision Tree is very simple to use because of its simple hierarchical flow. Therefore, it is mostly used for students' classification as compared to other data mining techniques. In Fig. 9, four data mining tools are reviewed that are used for students' exam performance prediction. New tools are rapidly emerging, however, mostly used tools are MATLAB, WEKA, RapidMiner and Python. WEKA is found as most frequently used tool in the present decade, followed by Python. WEKA is freely available software under a public license, but RapidMiner and MATLAB required to purchase a license. WEKA is easy to use software as it allows java code implementation as well as graphical user interface.

Above figures show that different algorithms can be used to predict students' results. All of the studies used different student' attributes as an input to their proposed prediction models. Mostly used attributes are demographics, attendance, academic results and students' clicks/views, students' personality, psychological factors and social behavior or activities. Figure 10 shows that students' academic records and demographic factors are proved as the best attributes in previous research studies. This survey paper also presents different feature selection techniques used to

Table 8 Data mining tools used for students' performance prediction

Year	MATLAB	Python	RapidMiner	WEKA
2003	-	-	-	(Yathongchai et al., 2003)
2005		-	-	(Bekele & Menzel, 2005)
2006	(Calvo-Flores et al., 2006)	-	-	(Al-Radaideh et al., 2006; Kalles & Pierrakeas, 2006)
2007	-	-	-	(Bresfelean, 2007; Nghe et al., 2007)
2009	(Lykourentzou et al., 2009)	-	-	(Nandeshwar and Chaudhari, 2009)
2011	(Baradwaj & Pal, 2012)	-	-	(Kabra & Bichkar, 2011)
2012	-	-	(Abu Tair & El-Halees, 2012)	(Osmanbegovic & Suljic, 2012; Ramaswami & Rathinasabapathy, 2012; Yadav & Pal, 2012)
2013	(Hidayah et al., 2013; Huang & Fang, 2013)	-	(Adhatrao et al., 2013)	(Kabakchieva, 2013; Ramesh et al., 2013; Romero et al., 2013; Sundar, 2013)
2014	(Mativo & Huang, 2014)	-	-	(Acharya & Sinha, 2014; Hu et al., 2014; Kulkarni & Ade, 2014; Osmanbegović et al., 2014; Pandey & Taruna, 2014)
2015	(Agrawal & Mavani, 2015; Iam-On & Boongoen, 2017)	(Brinton & Chiang, 2015)	-	(Ahmad et al., 2015; Anuradha & Velmurugan, 2015; Barbosa Manhães et al., 2015; Jishan et al., 2015; Kaur et al., 2015; Ruby & David, 2015)
2016	(Asogbon et al., 2016; Marbouti et al., 2016)	-	(Altujjar et al., 2016, Majeed & Junejo, 2016; Saa, 2016; Upadhyay & Gautam, 2016)	(Altujjar et al., 2016; Hamoud, 2016; Ilic et al., 2016; Kaur & Singh, 2016; Márquez-Vera et al., 2016; Mueen et al., 2016; Pandey & Taruna, 2016)
2017	(Abubakar & Ahmad, 2017)	(Rovira et al., 2017; Wang et al., 2017)	(Amra & Maghari, 2017; Asif et al., 2017; Thakar & Mehta, 2017; Wati et al., 2017)	(Costa et al., 2017; Daud et al., 2017; Khasanah, 2017; Senthil & Lin, 2017; Sultana et al., 2017)

Table 8 (continued)

Year	MATLAB	Python	RapidMiner	WEKA
2018	(Iyanda et al., 2018; Akinrotimi et al., 2018)	(Heuer & Breiter, 2018; Kaur & Bathla, 2018; Perez et al., 2018; Polyzou & Karypis, 2018; Deepika & Sathyanarayana, 2019)	(Adejo & Connolly, 2018; Miguéis et al., 2018)	(Shaziya et al., 2015; Al-Obeidat et al., 2018; Anoopkumar & Rahman, 2018; Borges et al., 2018; Hamoud et al., 2018; Helal et al., 2018; Hussain et al., 2018; Kiu, 2018; Livieris et al., 2018; Mishra & Kumawat, 2018; Oyefolahan et al., 2018; Pandey & Taruna, 2018; Puarungroj et al., 2018; Saheed et al., 2018; Wong & Senthil, 2018; Dey 2020)
2019	(Sari & Sunyoto, 2019; Turabieh, 2019; Yağci & Çevik, 2019)	(Bruce, 2019; Czibula et al., 2019; Mi, 2019; Ramaswami et al., 2019; Tripathi et al., 2019; Wu et al., 2019b, Xu et al., 2019; Zohair, 2019; Aydoğdu, 2020; Hew et al., 2020)	(Amazona & Hernandez, 2019; Kamal & Ahuja, 2019; Rojanavasu, 2019; Yaacob et al., 2019)	(Aman et al., 2019; Buenaño-Fernández et al., 2019; Francis & Babu, 2019; Hamoud & Humadi, 2019; Hasan, 2019; Meghji et al., 2019; Mikroskil, 2019; Pattanaphanchai et al., 2019; Salal et al., 2019; Umar, 2019)
2020	(Çevik & Tabaru-Örnek, 2020; Naicker et al., 2020; Sood & Saini, 2020; Tomasevic et al., 2020)	(Fachrie, 2019; Bhutto et al., 2020; Bravo et al., 2020; Freitas et al., 2020; Li et al., 2020; Moreno-Marcos et al., 2020; Mubarak et al., 2020; Pereira et al., 2020; Rajak et al., 2020; Yu et al., 2020; Zulfiker et al., 2020)	(Gil et al., 2020; Patacsil, 2020)	(Koutina & Kermanidis, 2011; Evwiekpaefe et al., 2014; Ahmed et al., 2020a; Ahmed et al., 2020c; Alhassan, 2020; Anwar & Rani, 2020; Mengash, 2020; Usman et al., 2020, Walia et al., 2020; Yahaya et al., 2020; Banu & Manjupargavi, 2021)

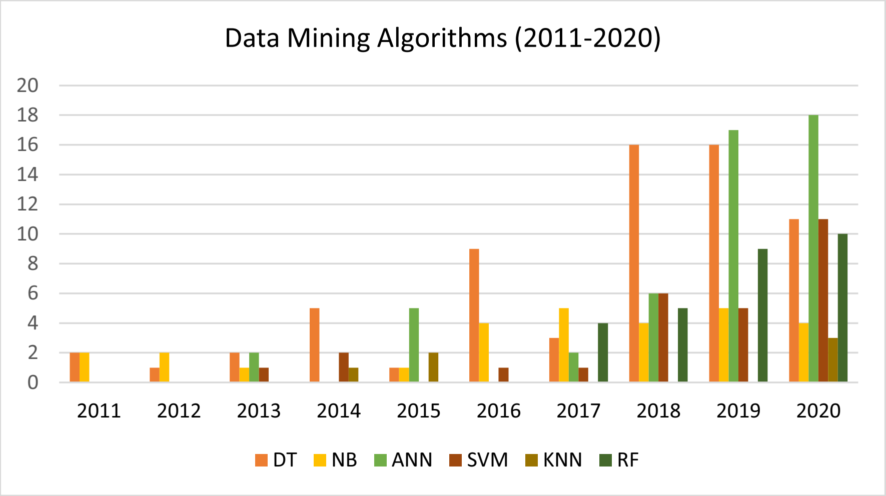

Fig. 7 Trending data mining algorithms used for students' performance prediction

select most influencing features. Figure 11 represents that more than a half of studies used feature selection methods before building prediction models. Having irrelevant features in the dataset may reduce the prediction results and increase model processing time. Figure 12 presents that feature selection methods are highly trending in past three years. Several feature selection methods are used in previous research paper; however, two techniques are widely used i.e., Information Gain and Gain Ratio.

7 Conclusion

Educational data mining gained a rapid growth as it helps institutions as well as policy makers in decision making. One of the most important research areas of educational data mining, is predicting students' future results based on their previous performance and demographics. Predicting future exam results before final examination can help teachers to find students who are at risk of failure, so they can be provided with extra assistance and time. Action plans can be implemented to prevent or reduce dropouts. This paper presents a summary of research studies conducted to predict students' performance using different data mining techniques. This study investigated recent twenty-years' work of researchers in order to compare different data mining techniques used for predictions and to evaluate students' attributes. Mostly used technique is Decision Tree, however, all data mining techniques gave different results because output of prediction models depends on the input data given to the model i.e., students' attributes. Mainly five types of students' attributes are used in the literature i.e., students' marks, attendance, learning behaviors, social activities and demographic data. It is found that students' marks or GPA is mostly used input type which gave best prediction

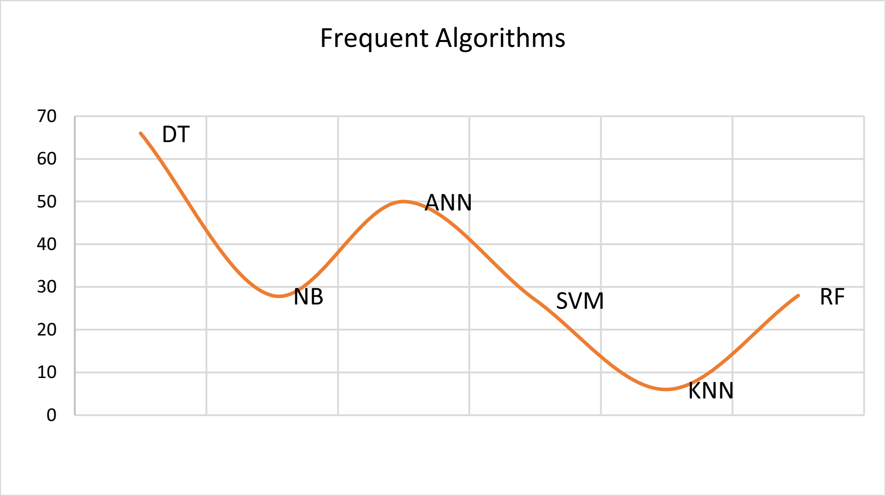

Fig. 8 Frequently used Data Mining Algorithms in last decade

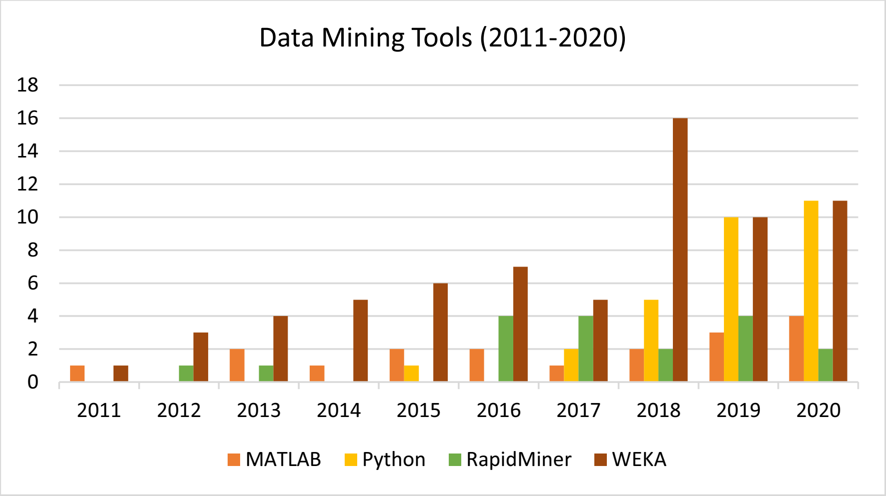

Fig. 9 Data mining tools used for students' performance prediction

results. The study also focused on data mining tools used for implementing data mining algorithms. Several data mining tools are available while four data mining tools are frequently used i.e., WEKA, MATLAB, Python, and RapidMiner. Several datasets are used in these research studies. Four mostly used datasets are OULAD, MOOCs, Moodle and UCI Repository Dataset.

The study shows that different evaluation methods including correlation, accuracy, f-measure, precision and recall are used. It is proved that all studies aim to classify students into binary i.e., pass/fail or multi classes i.e., grades. A few

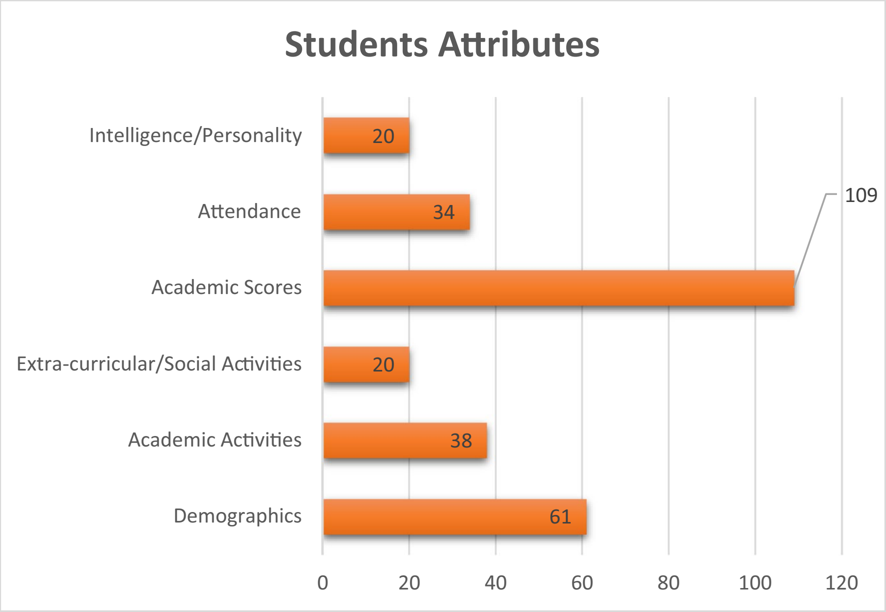

Fig. 10 Mostly used students' attributes

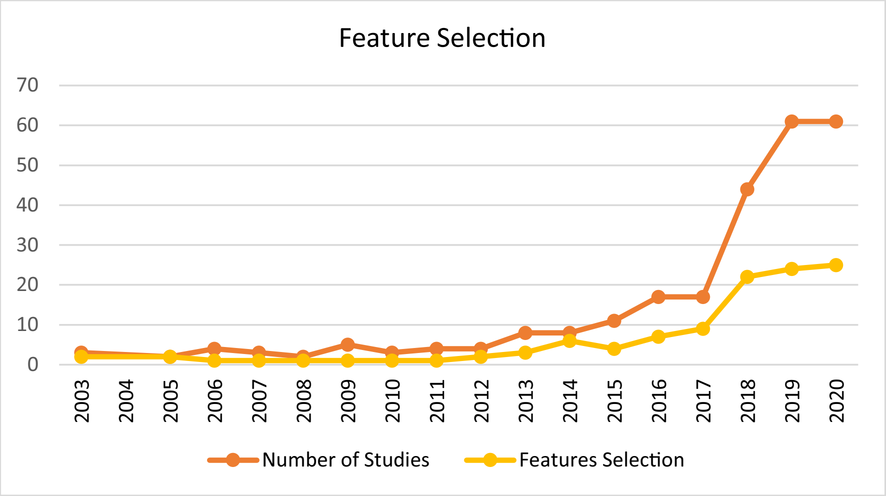

Fig. 11 Feature selection techniques

studies predict students' final marks or CGPA using regression techniques. The presented study also concluded that students' performance can be predicted at different stages e.g., at the time of admission, at the start of semester, and before final examinations. However, it is proved that prediction in the last two weeks of

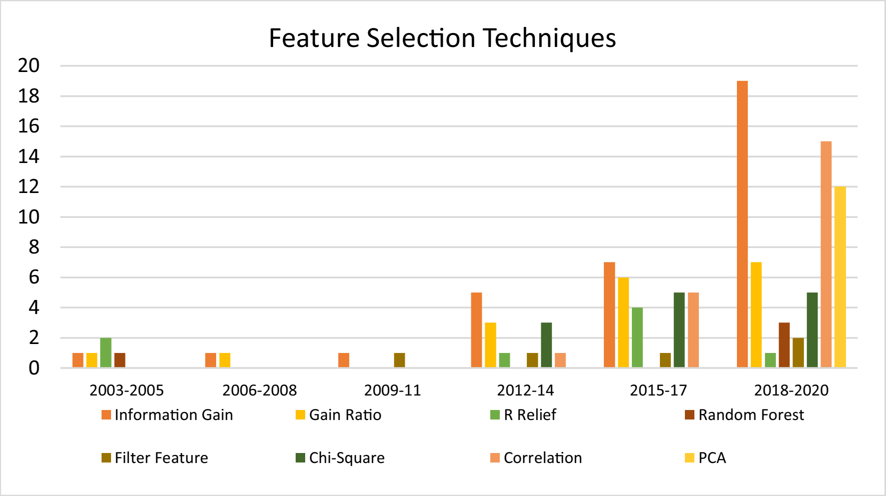

Fig. 12 Yearly trending feature selection techniques

semester can be more accurate as more academic features are available at this phase. Feature selection methods are trending in the past three years. Several studies proved that using only relevant features increases the prediction accuracy. This review will be beneficial for future research in predicting students' results and for institutions to pick the best classifier based on their students' data. This study will help academic policy makers and administrations to use their students' data in improving institutions' results, in available students' attributes and tools. The findings of the study will be helpful for future research studies to focus on highly influencing attributes only.

8 Limitations and future work

This study tried to provide a systematic review of research conducted to predict students' academic performance prediction. The number of research studies and algorithms explored are limited as each method cannot be mentioned in a single study. However, the survey provides a clear insight to effective and mostly used data mining algorithms, tools and students' attributes.

For future work, it is recommended to universities and online educational institutes using data mining for students' performance prediction and designing action plans to prevent students' dropout and increase courses' completion rates. Exploring students' psychological factors, teaching & learning methods, institutes' physical facilities and their impact on students' academic results is an open research area in EDM.

Acknowledgement: This research was partly supported by the Technology Development Program of MSS [No. S3033853] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A1031509).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: A case study. *International Journal of Information, 2*(2).

Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of students' performance in e-learning environment using random forest. *International Journal of Innovative Computing, 7*(2).

Acharya, A., & Sinha, D. (2014). Early prediction of students performance using machine learning techniques. *International Journal of Computer Applications, 107*(1).

Adebayo, A. O., & Chaubey, M. S. (2019). Data mining classification techniques on the analysis of student's performance. *GSJ, 7*(4), 45–52.

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education.*

Adekitan, A. I., & Noma-Osaghae, E. (2019). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies, 24*(2), 1527–1543.

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon, 5*(2), e01250.

Adekitan, A. I., & Salau, O. (2020). Toward an improved learning process: The relevance of ethnicity to data mining prediction of students' performance. *SN Applied Sciences, 2*(1), 1–15.

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting students' performance using ID3 and C4. In *5 classification algorithms* arXiv preprint arXiv:1310.2071.

Agrawal, H., & Mavani, H. (2015). Student performance prediction using machine learning. *International Journal of Engineering Research and Technology, 4*(03), 111–113.

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences, 9*(129), 6415–6426.

Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. *Bulletin of Education and Research, 40*(3), 157–164.

Ahmed, M. R., Tahid, S. T. I., Mitu, N. A., Kundu, P., & Yeasmin, S. (2020a). A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In *Paper presented at the 2020 11th international conference on computing, communication and networking technologies (ICCCNT).*

Ahmed, S. A., Billah, M. A., & Khan, S. I. (2020b). A machine learning approach to performance and dropout prediction in computer science: Bangladesh perspective. In *Paper presented at the 2020 11th international conference on computing, communication and networking technologies (ICCCNT).*

Ahmed, S. T., Al-Hamdani, R., & Croock, M. S. (2020c). Enhancement of student performance prediction using modified K-nearest neighbor. *Telkomnika, 18*(4), 1777–1783.

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2021). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences, 101102*.

Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments, 6*(1), 4.

Akinrotimi, A. O., Aremu, D. R., & Reuben, D. (2018). Student performance prediction using random student performance prediction using random tree and C4. 5 Algorithm ree and C4. 5 Algorithm.

Al-Obeidat, F., Tubaishat, A., Dillon, A., & Shah, B. (2018). Analyzing students' performance using multi-criteria classification. *Cluster Computing, 21*(1), 623–632.

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006). Mining student data using decision trees. In *Paper presented at the international Arab conference on information technology (ACIT'2006)*. Yarmouk University.

Alhassan, A., Zafar, B., & Mueen, A. (2020). Predict students' academic performance based on their assessment grades and online activity data. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(4).

Aljohani, N. R., Fayoumi, A., & Hassan, S.-U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability, 11*(24), 7238.

Alloghani, M., Al-Jumeily, D., Baker, T., Hussain, A., Mustafina, J., & Aljaaf, A. J. (2018). Applications of machine learning techniques for software engineering learning and early prediction of students' performance. In *Paper presented at the international conference on soft computing in data science*.

Altaf, S., Soomro, W., & Rawi, M. I. M. (2019). Student performance prediction using multi-layers artificial neural networks: A case study on educational data mining. In *Paper presented at the proceedings of the 2019 3rd international conference on information system and data mining*.

Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting critical courses affecting students performance: A case study. *Procedia Computer Science, 82*, 65–71.

Aluko, R. O., Adenuga, O. A., Kukoyi, P. O., Soyingbe, A. A., & Oyedeji, J. O. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: Using machine-learning techniques. *Construction Economics and Building, 16*(4), 86.

Aluko, R. O., Daniel, E. I., Oshodi, O. S., Aigbavboa, C. O., & Abisuga, A. O. (2018). Towards reliable prediction of academic performance of architecture students using data mining techniques. *Journal of Engineering, Design and Technology.*

Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M. (2019). A predictive model for predicting students academic performance. In *Paper presented at the 2019 10th international conference on information, intelligence, systems and applications (IISA).*

Amazona, M. V., & Hernandez, A. A. (2019). Modelling student performance using data mining techniques: Inputs for academic program development. In *Paper presented at the proceedings of the 2019 5th international conference on computing and data engineering*.

Amra, I. A. A., & Maghari, A. Y. (2017). Students performance prediction using KNN and Naïve Bayesian. In *Paper presented at the 2017 8th international conference on information technology (ICIT).*

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119–136.

Anoopkumar, M., & Rahman, A. (2018). Model of tuned J48 classification and analysis of performance prediction in educational data mining. *The International Journal of Applied Engineering Research (IJAER), 13*(20), 14717–14727.

Anuradha, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian Journal of Science and Technology, 8*(15), 1–12.

Anwar, M. A., & Rani, R. (2020). Data science for prediction of grades in a mathematics course based on performance in its prerequisites.

Arsad, P. M., & Buniyamin, N. (2013). A neural network students' performance prediction model (NNSPPM). In *Paper presented at the 2013 IEEE international conference on smart instrumentation, measurement and applications (ICSIMA).*

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177–194.

Asogbon, M. G., Samuel, O. W., Omisore, M. O., & Ojokoh, B. A. (2016). *A multi-class support vector machine approach for students academic performance prediction* (p. 4).

Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. *Education and Information Technologies, 25*(3), 1913–1927.

Banu, S. R., & Manjupargavi, R. (2021). Performance analysis and prediction of students results using machine learning and big data approach. In *Paper presented at the in 2021 2nd international conference on smart electronics and communication (ICOSEC).*

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv*, 1201.3417.

Barbosa Manhães, L. M., da Cruz, S. M. S., & Zimbrão, G. (2015). Towards automatic prediction of student performance in STEM undergraduate degree programs. In *Paper presented at the proceedings of the 30th annual ACM symposium on applied computing*.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Mahmood, T., & Hussain, A. (2021). A random forest students' performance prediction (rfspp) model based on students' demographic features. In *Paper presented at the 2021 Mohammad Ali Jinnah University international conference on computing (MAJICC).*

Bekele, R., & McPherson, M. (2011). A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition. *British Journal of Educational Technology, 42*(3), 395–416.

Bekele, R., & Menzel, W. (2005). A bayesian approach to predict performance of a student (bapps): A case with ethiopian students. *Algorithms, 22*(23), 24.

Bhardwaj, B. K., & Pal, S. (2012). Data mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418.

Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020). Predicting students' academic performance through supervised machine learning. In *Paper presented at the 2020 international conference on information science and communication technology (ICISCT).*

Borges, V. R. P., Esteves, S., de Nardi Araújo, P., de Oliveira, L. C., & Holanda, M. (2018). Using principal component analysis to support students' performance prediction and data analysis. In *Paper presented at the Brazilian symposium on computers in education (Simpósio Brasileiro de Informática na Educação-SBIE).*

Bravo, L. E. C., Molano, J. I. R., & Trujillo, E. R. (2020). Exploration of a system to determine the academic performance of engineering students through machine learning. *International Journal of Advanced Science and Technology, 29*(7), 11894–11905.

Bresfelean, V. P. (2007). Analysis and predictions on students' behavior using decision trees in Weka environment. In *Paper presented at the 2007 29th international conference on information technology interfaces*.

Brinton, C. G., & Chiang, M. (2015). MOOC performance prediction via clickstream data and social learning networks. In *Paper presented at the 2015 IEEE conference on computer communications (INFOCOM).*

Bruce, A. (2019). The prediction of student performance through the use of machine learning.

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability, 11*(10), 2833.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering, 66*, 541–556.

Burman, I., & Som, S. (2019). Predicting students academic performance using support vector machine. In *Paper presented at the 2019 Amity International conference on artificial intelligence (AICAI).*

Calvo-Flores, M. D., Galindo, E. G., Jiménez, M. P., & Pineiro, O. P. (2006). Predicting students' marks from Moodle logs using neural network models. *Current Developments in Technology-Assisted Education, 1*(2), 586–590.

Cao, Y., Gao, J., Lian, D., Rong, Z., Shi, J., Wang, Q., & Zhou, T. (2018). Orderliness predicts academic performance: Behavioural analysis on campus lifestyle. *Journal of the Royal Society Interface, 15*(146), 20180210.

Cavazos, R., & Garza, S. E. (2017). Learning models for student performance prediction. *Paper presented at the Mexican International Conference on Artificial* Intelligence.

Çevik, M., & Tabaru-Örnek, G. (2020). Comparison of MATLAB and SPSS software in the prediction of academic achievement with artificial neural networks: Modeling for elementary school students. *International Online Journal of Education and Teaching, 7*(4), 1689–1707.

Chand, K. S. P., Prabakaran, N., Ramani, S., Rao, D. V., & Vemparala, S. (2020). Assessment analysis and performance prediction using M5 rules.

Chang, C.-T., Tu, C.-S., & Hajiyev, J. (2019). Integrating academic type of social media activity with perceived academic performance: A role of task-related and non-task-related compulsive internet use. *Computers & Education, 139*, 157–172.

Chanlekha, H., & Niramitranon, J. (2018). Student performance prediction model for early-identification of at-risk students in traditional classroom settings. In *Paper presented at the proceedings of the 10th international conference on Management of Digital EcoSystems*.

Chen, Y., Zheng, Q., Ji, S., Tian, F., Zhu, H., & Liu, M. (2020). Identifying at-risk students based on the phased prediction model. *Knowledge and Information Systems, 62*(3), 987–1003.

Chounta, I.-A., & Carvalho, P. F. (2019). Square it up! How to model step duration when predicting student performance. In *Paper presented at the proceedings of the 9th international conference on Learning Analytics & Knowledge*.

Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior, 107*, 105584.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256.

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems, 135*, 113325.

Czibula, G., Mihai, A., & Crivei, L. M. (2019). S PRAR: A novel relational association rule mining classification model applied for academic performance prediction. *Procedia Computer Science, 159*, 20–29.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Paper presented at the proceedings of the 26th international conference on world wide web companion*.

Deepika, K., & Sathyanarayana, N. (2019). Relief-F and budget tree random forest based feature selection for student academic performance prediction. *International Journal of Intelligent Engineering and Systems, 12*(1), 30–39.

Devasia, T., Vinushree, T., & Hegde, V. (2016). Prediction of students performance using educational data mining. In *Paper presented at the 2016 international conference on data mining and advanced computing (SAPIENCE)*.

DEY, A. (2020). *Prediction and analysis of student performance by data mining in WEKA*. West Bengal University of Technology.

Evwiekpaefe, A. E., Isa, M. M., & Ajakaiye, F. (2014). Analyzing factors affecting academic performance of postgraduate students using data mining techniques.

Fachrie, M. (2019). Development of educational data mining model for predicting student punctuality and graduation predicate. *International Journal of Technology and Engineering Studies, 5*(5), 151–156.

Farissi, A., & Dahlan, H. M. (2020). Genetic algorithm based feature selection with ensemble methods for student academic performance prediction. In *Paper presented at the journal of physics: Conference series*.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research, 94*, 335–343.

Figueroa-Cañas, J., & Sancho-Vinuesa, T. (2020). Early prediction of dropout and final exam performance in an online statistics course. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 15*(2), 86–94.

Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of Medical Systems, 43*(6), 1–15.

Freitas, F., Vasconcelos, F., Peixoto, F., Hassan, M., Ali Akber Dewan, M., & de Albuquerque, V. (2020). IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics, 9*(10), 1613.

Gil, P. D., da Cruz Martins, S., Moro, S., & Costa, J. M. (2020). A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, 1–26.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In *Paper presented at the 2015 international symposium on educational technology (ISET)*.

Hamoud, A. (2016). Selection of best decision tree algorithm for prediction and classification of students' action. *American International Journal of Research in Science, Technology, Engineering & Mathematics, 16*(1), 26–32.

Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence, 5*, 26–31.

Hamoud, A., & Humadi, A. (2019). Student's success prediction model based on artificial neural networks (ANN) and a combination of feature selection methods. *Journal of Southwest Jiaotong University, 54*(3).

Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology, 25*, 326–332.

Harvey, J. L., & Kumar, S. A. (2019). A practical model for educators to predict student performance in K-12 education using machine learning. In *Paper presented at the 2019 IEEE symposium series on computational intelligence (SSCI).*

Hasan, H. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019). Machine learning algorithm for student's performance prediction. In *Paper presented at the 2019 10th international conference on computing, communication and networking technologies (ICCCNT).*

Hasan, M. (2019). *Predicting student performance to reduce dropout using J48 decision tree algorithm*. Daffodil International University.

He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., & Jiang, B. (2020). Online at-risk student identification using RNN-GRU joint neural networks. *Information, 11*(10), 474.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems, 161*, 134–146.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks Vis-à-Vis regression. *New Directions for Institutional Research, 2006*(131), 17–33.

Heuer, H., & Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. In *DeLFI 2018-Die 16*. Fachtagung Informatik.

Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education, 145*, 103724.

Hidayah, I., Permanasari, A. E., & Ratwastuti, N. (2013). Student classification for academic performance prediction using neuro fuzzy in a conventional classroom. In *Paper presented at the 2013 international conference on information technology and electrical engineering (ICITEE).*

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education, 37*, 66–75.

Hsu, P.-L., Lai, R., & Chiu, C. (2003). The hybrid of association rule algorithms and genetic algorithms for tree induction: An example of predicting the student course performance. *Expert Systems with Applications, 25*(1), 51–62.

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, 36*, 469–478.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education, 61*, 133–145.

Hughes, G., & Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). *Research and Practice in Technology Enhanced Learning, 10*(1), 1–18.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science, 9*(2), 447–459.

Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodorou, P., Kurtoglu, F., & Hazarika, G. (2019). Prediction model on student performance based on internal assessment using deep learning. *iJET, 14*(8), 4–22.

Iam-On, N., & Boongoen, T. (2017). Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *International Journal of Machine Learning and Cybernetics, 8*(2), 497–510.

Ilic, M., Spalevic, P., Veinovic, M., & Alatresh, W. S. (2016). Students' success prediction using Weka tool. *Infoteh-Jahorina, 15*, 684–688.

Iyanda, A. R., Ninan, O. D., Ajayi, A. O., & Anyabolu, O. G. (2018). Predicting student academic performance in computer science courses: A comparison of neural network models. *International Journal of Modern Education & Computer Science, 10*(6).

Jha, N. I., Ghergulescu, I., & Moldovan, A.-N. (2019). OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques. In *Paper presented at the CSEDU (2)*.

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority oversampling technique. *Decision Analytics, 2*(1), 1–25.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies, 13*(1), 61–72.

Kabra, R., & Bichkar, R. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications, 36*(11), 8–12.

Kalles, D., & Pierrakeas, C. (2006). Analyzing student performance in distance learning with genetic algorithms and decision trees. *Applied Artificial Intelligence, 20*(8), 655–674.

Kamal, P., & Ahuja, S. (2019). *An ensemble-based model for prediction of academic performance of students in undergrad professional course*. Journal of Engineering.

Karamouzis, S. T., & Vrettos, A. (2008). An artificial neural network for predicting student graduation outcomes. In *Paper presented at the proceedings of the world congress on engineering and computer science*.

Karimi, H., Derr, T., Huang, J., & Tang, J. (2020). Online academic course performance prediction using relational graph convolutional neural network. In *Paper presented at the proceedings of the 13th international conference on educational data mining (EDM 2020)*.

Karlık, M., & Karlık, B. (2020). Prediction of student's performance with deep neural networks. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*.

Kaur, G., & Singh, W. (2016). Prediction of student performance using weka tool. *An International Journal of Engineering Sciences, 17*, 8–16.

Kaur, H., & Bathla, E. G. (2018). Student performance prediction using educational data mining techniques. International journal on future revolution in Computer Science & Communication. *Engineering, 4*(12), 93–97-93–97.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science, 57*, 500–508.

Kaviyarasi, R., & Balasubramanian, T. (2018). Exploring the high potential factors that affects students' academic performance. *International Journal of Education and Management Engineering, 8*(6), 15–23.

Kaviyarasi, R., & Balasubramanian, T. (2020). Predictive analysis of academic performance of college students using ensemble stacking. *Kongunadu Research Journal, 7*(2), 94–98.

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10*(1), 28–47.

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies, 26*(1), 205–240.

Khasanah, A. U. (2017). A comparative study to predict student's performance using educational data mining techniques. In *Paper presented at the IOP conference series: Materials science and engineering*.

Khazaaleh, M. K. (2020). Predictive model to predict the test scores of the computer skills-2 course for future students .

Kiu, C.-C. (2018). Data mining analysis on student's academic performance through exploration of student's background and social activities. In *Paper presented at the 2018 fourth international conference on advances in computing, Communication & Automation (ICACCA)*.

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems, 23*(6), 529–535.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2002). *Efficiency of machine learning techniques in predicting students' performance in distance learning systems*. University of Patras, Greece.

Kotsiantis, S. B., & Pintelas, P. E. (2005). Predicting students marks in hellenic open university. In *Paper presented at the fifth IEEE international conference on advanced learning technologies (ICALT'05)*.

Koutina, M., & Kermanidis, K. L. (2011). Predicting postgraduate students' performance using machine learning techniques. In *Artificial intelligence applications and innovations* (pp. 159–168). Springer.

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

Kulkarni, P., & Ade, R. (2014). Prediction of student's performance based on incremental learning. *International Journal of Computer Applications, 99*(14), 10–16.

Kumar, T. R., Vamsidhar, T., Harika, B., Kumar, T. M., & Nissy, R. (2019). Students performance prediction using data mining techniques. In *Paper presented at the 2019 international conference on intelligent sustainable systems (ICISS).*

Kumar, V., & Garg, M. (2019). Comparison of machine learning models in student result prediction. In *Paper presented at the international conference on advanced computing networking and informatics*.

Lenin, T., & Chandrasekaran, N. (2019). Students' performance prediction Modelling using classification technique in R.

Li, J., Sun, S., Yin, H., Dawson, P., & Doss, R. (2020). SEPN: A sequential engagement based academic performance prediction model. *IEEE Intelligent Systems.*

Liang, J., Li, C., & Zheng, L. (2016). Machine learning application in MOOCs: Dropout prediction. In *Paper presented at the 2016 11th international conference on Computer Science & Education (ICCSE).*

Lin, J., Imbrie, P., & Reid, K. J. (2009). Student retention modelling: An evaluation of different methods and their impact on prediction results. *Research in Engineering Education Sysmposium*, 1–6.

Liu, H., Zhu, Y., Zang, T., Yu, J., & Cai, H. (2020). Jointly modeling individual student behaviors and social influence for prediction tasks. In *Paper presented at the proceedings of the 29th ACM international conference on Information & Knowledge Management.*

Liu, W. (2019). An improved back-propagation neural network for the prediction of college students' english performance. *International Journal of Emerging Technologies in Learning, 14*(16).

Livieris, I. E., Drakopoulou, K., Mikropoulos, T. A., Tampakas, V., & Pintelas, P. (2018). An ensemble-based semi-supervised approach for predicting students' performance. In *Research on e-learning and ICT in education* (pp. 25–42). Springer.

Lottering, R., Hans, R., & Lall, M. (2020). A model for the identification of students at risk of dropout at a university of technology. In *Paper presented at the 2020 international conference on artificial intelligence, big data, computing and data communication systems (icABCD).*

Lu, H., & Yuan, J. (2018). Student performance prediction model based on discriminative feature selection. *International Journal of Emerging Technologies in Learning, 13*(10).

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology, 60*(2), 372–380.

Majeed, E. A., & Junejo, K. N. (2016). Grade prediction using supervised machine learning techniques. In *E-proceedings of the 4th global summit on education*.

Makombe, F., & Lall, M. (2020). A predictive model for the determination of academic performance in private higher education institutions. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(9).

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103*, 1–15.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems, 33*(1), 107–124.

Masood, M. F., Khan, A., Hussain, F., Shaukat, A., Zeb, B., & Ullah, R. M. K. (2019). Towards the selection of best machine learning model for student performance analysis and prediction. In *Paper presented at the 2019 6th international conference on Soft Computing & Machine Intelligence (ISCMI).*

Mativo, J. M., & Huang, S. (2014). Prediction of students' academic performance: Adapt a methodology of predictive modeling for a small sample size. In *Paper presented at the 2014 IEEE Frontiers in education conference (FIE) proceedings*.

Meghji, A. F., Mahoto, N. A., Unar, M. A., & Shaikh, M. A. (2019). Predicting student academic performance using data generated in higher educational institutes.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access, 8*, 55462–55470.

Mi, C. (2019). Data-driven student learning performance prediction based on RBF neural network. *International Journal of Performability Engineering, 15*(6), 1560.

Mi, C., Peng, X., Cai, Z., Deng, Q., & Zhao, C. (2018). A genetic algorithm based method of early warning rule mining for student performance prediction. In *Paper presented at the international conference on cloud computing and security*.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51.

Mikroskil, S. (2019). Information systems students' study performance prediction using data mining approach.

Mishra, T., & Kumawat, C. (2018). Critical evaluation of classification algorithms for performance prediction in higher education setup.

Moreno-Marcos, P. M., Pong, T.-C., Munoz-Merino, P. J., & Kloos, C. D. (2020). Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access, 8*, 5264–5282.

Moseley, L. G., & Mead, D. M. (2008). Predicting who will drop out of nursing courses: A machine learning exercise. *Nurse Education Today, 28*(4), 469–475.

Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 1–20.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education & Computer Science, 8*(11).

Mutanu, L., & Machoka, P. (2019). Enhancing computer students' academic performance through predictive modelling-a proactive approach. In *Paper presented at the 2019 14th international conference on Computer Science & Education (ICCSE)*.

Nabizadeh, S., Hajian, S., Sheikhan, Z., & Rafiei, F. (2019). Prediction of academic achievement based on learning strategies and outcome expectations among medical students. *BMC Medical Education, 19*(1), 99.

Naicker, N., Adeliyi, T., & Wing, J. (2020). *Linear support vector machines for prediction of student performance in school-based education. mathematical problems in engineering*, 2020.

Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237.

Nandeshwar, A., & Chaudhari, S. (2009). Enrollment prediction models using data mining. Retrieved January, 10, 2010.

Narayanasamy, S. K., & Elçi, A. (2020). An effective prediction model for online course dropout rate. *International Journal of Distance Education Technologies (IJDET), 18*(4), 94–110.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Paper presented at the 2007 37th annual frontiers in education conference-global engineering: Knowledge without borders, opportunities without passports*.

Nuankaew, W., & Thongkam, J. (2020). Improving student academic performance prediction models using feature selection. In *Paper presented at the 2020 17th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*.

Ogor, E. N. (2007). Student academic performance monitoring and evaluation using data mining techniques. In *Paper presented at the electronics, robotics and automotive mechanics conference (CERMA 2007)*.

Olalekan, A. M., Egwuche, O. S., & Olatunji, S. O. (2020). Performance evaluation of machine learning techniques for prediction of graduating students in tertiary institution. In *Paper presented at the 2020 international conference in mathematics, computer engineering and computer science (ICMCECS)*.

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business, 10*(1), 3–12.

Osmanbegović, E., Suljić, M., & Agić, H. (2014). Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija, 16*(34), 147–158.

Oyefolahan, I. O., Idris, S., Etuk, S. O., & Alabi, I. O. (2018). Academic performance prediction for success rate improvement in higher institutions of learning: An application of data mining classification algorithms.

Paliwal, M., & Kumar, U. A. (2009). A study of academic performance of business school graduates using neural network and statistical techniques. *Expert Systems with Applications, 36*(4), 7865–7872.

Pandey, M., & Taruna, S. (2014). A multi-level classification model pertaining to the student's academic performance prediction. *International Journal of Advances in Engineering & Technology, 7*(4), 1329.

Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science, 8*, 364–366.

Pandey, M., & Taruna, S. (2018). An ensemble-based decision support system for the students' academic performance prediction. In *ICT Based Innovations* (pp. 163–169). Springer.

Patacsil, F. F. (2020). Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. *Universal Journal of Educational Research, 8*(9), 4036–4047.

Patil, P. A., & Mane, R. (2014). Prediction of students performance using frequent pattern tree. In *Paper presented at the 2014 international conference on computational intelligence and communication networks*.

Patil, R., Salunke, S., Kalbhor, M., & Lomte, R. (2018). Prediction system for student performance using data mining classification. In *Paper presented at the 2018 fourth international conference on computing communication control and automation (ICCUBEA)*.

Pattanaphanchai, J., Leelertpanyakul, K., & Theppalak, N. (2019). The investigation of student dropout prediction model in thai higher education using educational data mining: A case study of faculty of science, prince of Songkla Uni-versity. *Journal of University of Babylon for Pure and Applied Sciences, 27*(1), 356–367.

Pereira, F. D., Fonseca, S. C., Oliveira, E. H., Oliveira, D. B., Cristea, A. I., & Carvalho, L. S. (2020). Deep learning for early performance prediction of introductory programming students: A comparative and explanatory study. *Brazilian Journal of Computers in Education, 28*, 723–749.

Pereira, F. D., Oliveira, E. H., Fernandes, D., & Cristea, A. (2019). Early performance prediction for CS1 course students using a combination of machine learning and an evolutionary algorithm. In *Paper presented at the 2019 IEEE 19th international conference on advanced learning technologies (ICALT)*.

Perez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. In *Paper presented at the 2018 IEEE 1st Colombian conference on applications in computational intelligence (CoLCACI)*.

Polyzou, A., & Karypis, G. (2018). *Feature extraction for classifying students based on their academic performance*. International Educational Data Mining Society.

Poudyal, S., Nagahi, M., Nagahisarchoghaei, M., & Ghanbari, G. (2020). Machine learning techniques for determining students' academic performance: A sustainable development case for engineering education. In *Paper presented at the 2020 international conference on decision aid sciences and application (DASA)*.

Puarungroj, W., Boonsirisumpun, N., Pongpatrakant, P., & Phromkhot, S. (2018). Application of data mining techniques for predicting student success in English exit exam. In *Paper presented at the proceedings of the 12th international conference on ubiquitous information management and communication*.

Qian, R., Sengan, S., & Juneja, S. (2022). *English language teaching based on big data analytics in augmentative and alternative communication system* (pp. 1–12).

Qu, S., Li, K., Fan, Z., Wu, S., Liu, X., & Huang, Z. (2019). Behavior pattern and compiled information based performance prediction in MOOCs. arXiv preprint arXiv:1908.01304.

Raga, R. C., & Raga, J. D. (2019). Early prediction of student performance in blended learning courses using deep neural networks. In *Paper presented at the 2019 international symposium on educational technology (ISET)*.

Rahman, M. H., & Islam, M. R. (2017). Predict student's academic performance and evaluate the impact of different attributes on the performance using data mining techniques. In *Paper presented at the 2017 2nd international conference on electrical & electronic engineering (ICEEE)*.

Rajak, A., Shrivastava, A. K., & Vidushi. (2020). Applying and comparing machine learning classification algorithms for predicting the results of students. *Journal of Discrete Mathematical Sciences and Cryptography, 23*(2), 419–427.

Ramaswami, G., Susnjak, T., Mathrani, A., Lim, J., & Garcia, P. (2019). *Using educational data mining techniques to increase the prediction accuracy of student academic performance*. Information and Learning Sciences.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv:1002.1144.

Ramaswami, M., & Rathinasabapathy, R. (2012). Student performance prediction. *International Journal of Computational Intelligence and Informatics, 1*(4), 231–235.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. *International Journal of Computer Applications, 63*(8).

Rifat, M. R. I., Al Imran, A., & Badrudduza, A. (2019). Educational performance analytics of undergraduate business students. *International Journal of Modern Education and Computer Science, 11*(7), 44.

Rincón-Flores, E. G., López-Camacho, E., Mena, J., & López, O. O. (2020). Predicting academic performance with artificial intelligence (AI), a new tool for teachers and students. In *Paper presented at the 2020 IEEE global engineering education conference (EDUCON)*.

Rojanavasu, P. (2019). Educational data analytics using association rule mining and classification. In *Paper presented at the 2019 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT-NCON)*.

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458–472.

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS One, 12*(2), e0171207.

Ruby, J., & David, K. (2015). Analysis of influencing factors in predicting students performance using MLP-A comparative study. *International Journal of Innovative Research in Computer and Communication Engineering, 3*(2), 1085–1092.

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications, 7*(5), 212–220.

Sahebi, S., & Brusilovsky, P. (2018). *Student performance prediction by discovering inter-activity relations*. International Educational Data Mining Society.

Saheed, Y., Oladele, T., Akanni, A., & Ibrahim, W. (2018). Student performance prediction based on data mining classification techniques. *Nigerian Journal of Technology, 37*(4), 1087–1091.

Saifudin, A., & Desyani, T. (2020). Forward selection technique to choose the best features in prediction of student academic performance based on naïve bayes. In *Paper presented at the journal of physics: Conference series*.

Salal, Y., Abdullaev, S., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *The International Journal of Engineering and Advanced Technology, 8*(4C), 54–59.

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. *The Internet and Higher Education, 37*, 76–89.

Santoso, H. B. (2020). Fuzzy decision tree to predict student success in their studies. *International Journal of Quantitative Research and Modeling, 1*(3), 135–144.

Sari, E. Y., & Sunyoto, A. (2019). Optimization of weight backpropagation with particle swarm optimization for student dropout prediction. In *Paper presented at the 2019 4th international conference on information technology, information systems and electrical engineering (ICITISEE)*.

Sawant, T. U., Pol, U. R., & Patankar, P. S. (2019). Educational data mining prediction model using decision tree algorithm. *International Journal of Emerging Technologies and Innovative Research, 2349*(5162), 306–313. www.jetir.org

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics* (pp. 99–111). Springer.

Senthil, S., & Lin, W. M. (2017). Applying classification techniques to predict students' academic results. In *Paper presented at the 2017 IEEE international conference on current trends in advanced computing (ICCTAC)*.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414–422.

Shaziya, H., Zaheer, R., & Kavitha, G. (2015). Prediction of students performance in semester exams using a Naïve Bayes classifier. *International Journal of Innovative Research in Science, Engineering and Technology, 4*(10), 9823–9829.

Singhani, S., Desai, S., Bailurkar, R., & Mantri, R. (2019). Student academic performance prediction using machine learning.

Sivakumar, S., & Selvaraj, R. (2018). Predictive modeling of students performance through the enhanced decision tree. In *Advances in electronics, communication and computing* (pp. 21–36). Springer.

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology, 9*(4), 1–5.

Sokkhey, P., & Okazaki, T. (2019). Comparative study of prediction models on high school student performance in mathematics. In *Paper presented at the 2019 34th international technical conference on circuits/systems, computers and communications (ITC-CSCC).*

Sokkhey, P., & Okazaki, T. (2020a). Developing web-based support systems for predicting poor-performing students using educational data mining techniques. *studies, 11*(7).

Sokkhey, P., & Okazaki, T. (2020b). Development and optimization of deep belief networks applied for academic performance prediction with larger datasets. *IEIE Transactions on Smart Processing & Computing, 9*(4), 298–311.

Sokkhey, P., & Okazaki, T. (2020c). Hybrid machine learning algorithms for predicting academic performance. *International Journal of Advanced Computer Science and Applications, 11*, 32–41.

Solís, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. In *Paper presented at the 2018 IEEE international work conference on bioinspired intelligence (IWOBI).*

Soni, A., Kumar, V., Kaur, R., & Hemavath, D. (2018). Predicting student performance using data mining techniques. *International Journal of Pure and Applied Mathematics, 119*(12), 221–227.

Sood, S., & Saini, M. (2020). Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation. *Education and Information Technologies*, 1–16.

Stančin, I., & Jović, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. In *Paper presented at the 2019 42nd international convention on information and communication technology, electronics and microelectronics (MIPRO).*

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., & Hu, G. (2018). Exercise-enhanced sequential modeling for student performance prediction. In *Paper presented at the proceedings of the AAAI conference on artificial intelligence*.

Sudha, M., & Kumaravel, A. (2017). Students'performance prediction based on rough sets. *Indian Journal of Computer Science and Engineering, 8*, 584–589.

Sukhbaatar, O., Usagawa, T., & Choimaa, L. (2019). An artificial neural network based early prediction of failure-prone students in blended learning course. *International Journal of Emerging Technologies in Learning (iJET), 14*(19), 77–92.

Sultana, S., Khan, S., & Abbas, M. A. (2017). Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering Education, 54*(2), 105–118.

Sundar, P. P. (2013). A comparative study for predicting students academic performance using Bayesian network classifiers. *IOSR Journal of Engineering (IOSRJEN)*, e-ISSN, 2250-3021.

Sweeney, M., Lester, J., & Rangwala, H. (2015). Next-term student grade prediction. In *Paper presented at the 2015 IEEE international conference on big data (big data).*

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research, 54*, 207–226.

Thakar, P., & Mehta, A. (2017). A unified model of clustering and classification to improve students' employability prediction. *International Journal of Intelligent Systems and Applications, 9*(9), 10.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education, 143*, 103676.

Tripathi, A., Yadav, S., & Rajan, R. (2019). Naive Bayes classification model for the student performance prediction. In *Paper presented at the 2019 2nd international conference on intelligent computing, instrumentation and control technologies (ICICICT).*

Turabieh, H. (2019). Hybrid machine learning classifiers to predict student performance. In *Paper presented at the 2019 2nd international conference on new trends in computing sciences (ICTCS).*

Ulloa-Cazarez, R. L., López-Martín, C., Abran, A., & Yáñez-Márquez, C. (2018). Prediction of online students performance by means of genetic programming. *Applied Artificial Intelligence, 32*(9–10), 858–881.

Umar, M. A. (2019). Student academic performance prediction using artificial neural networks: A case study. *International Journal of Computer Applications, 975*, 8887.

Upadhyay, H., Juneja, S., Juneja, A., Dhiman, G., & Kautish, S. (2021). Evaluation of ergonomics-related disorders in online education using fuzzy AHP. *Computational Intelligence and Neuroscience*, 2021.

Upadhyay, J., & Gautam, P. (2016). Effect of numerous data sets on performance prediction. *International Journal of Computer Applications, 147*(5).

Usman, M. M., Owolabi, O., & Ajibola, A. A. (2020). Feature selection: It importance in performance prediction.

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of predicting student's performance using machine learning algorithms. *International Journal of Intelligent Systems and Applications, 11*(12), 34.

Vital, T. P., Sangeeta, K., & Kumar, K. K. (2021). Student classification based on cognitive abilities and predicting learning performances using machine learning models. *International Journal of Computing and Digital Systems, 10*(1), 63–75.

Vivek Raj, S., & Manivannan, S. (2020). Predicting student failure in university examination using machine learning algorithms. *forest, 84*(66.14), 0.24.

Vora, D. R., & Rajamani, K. (2019). A hybrid classification model for prediction of academic performance of students: A big data application. *Evolutionary Intelligence*, 1–14.

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior, 104*, 106189.

Wakelam, E., Jefferies, A., Davey, N., & Sun, Y. (2020). The potential for student performance prediction in small cohorts with minimal available attributes. *British Journal of Educational Technology, 51*(2), 347–370.

Walia, N., Kumar, M., Nayar, N., & Mehta, G. (2020). Student's academic performance prediction in academic using data mining techniques. *Available at SSRN, 3565874*.

Wang, W., Yu, H., & Miao, C. (2017). Deep model for dropout prediction in MOOCs. In *Paper presented at the proceedings of the 2nd international conference on crowd science and engineering*.

Wati, M., Indrawan, W., Widians, J. A., & Puspitasari, N. (2017). Data mining for predicting students' learning result. In *Paper presented at the 2017 4th international conference on computer applications and information processing technology (CAIPT)*.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). Delving deeper into MOOC student dropout prediction. *arXiv preprint arXiv*, 1702.06404.

Wiyono, S., Wibowo, D. S., Hidayatullah, M. F., & Dairoh, D. (2020). Comparative study of KNN, SVM and decision tree algorithm for student's performance prediction. *IJCSAM (International Journal of Computing Science and Applied Mathematics), 6*(2), 50–53.

Wong, J. C. F., & Yip, T. C. Y. (2020). Measuring students' academic performance through educational data mining. *International Journal of Information and Education Technology, 10*(11).

Wong, M. L., & Senthil, S. (2018). Applying attribute selection algorithms in academic performance prediction. In *Paper presented at the international conference on intelligent data communication technologies and internet of things*.

Wook, M., Yahaya, Y. H., Wahab, N., Isa, M. R. M., Awang, N. F., & Seong, H. Y. (2009). Predicting NDUM student's academic performance using data mining techniques. In *Paper presented at the 2009 second international conference on computer and electrical engineering*.

Wu, B., Qu, S., Ni, Y., Zhou, Y., Wang, P., & Li, Q. (2019a). Predicting student performance using weblogs. In *Paper presented at the 2019 14th international conference on Computer Science & Education (ICCSE)*.

Wu, N., Zhang, L., Gao, Y., Zhang, M., Sun, X., & Feng, J. (2019b). CLMS-Net: Dropout prediction in MOOCs with deep learning. In *Paper presented at the proceedings of the ACM Turing Celebration Conference-China*.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior, 98*, 166–173.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indonesian Journal of Electrical Engineering and Computer Science, 16*(3), 1584–1592.

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv*, 1203.3832.

Yağci, A., & Çevik, M. (2019). Prediction of academic achievements of vocational and technical high school (VTS) students in science courses through artificial neural networks (comparison of Turkey and Malaysia). *Education and Information Technologies, 24*(5), 2741–2761.

Yahaya, C. A. C., Yaakub, C. Y., Abidin, A. F. Z., Ab Razak, M. F., Hasbullah, N. F., & Zolkipli, M. F. (2020). The prediction of undergraduate student performance in chemistry course using multilayer perceptron. In *Paper presented at the IOP conference series: Materials science and engineering*.

Yathongchai, W., Yathongchai, C., Kerdprasop, K., & Kerdprasop, N. (2003). Factor analysis with data mining technique in higher educational student drop out. *Latest Advances in Educational Technologies*.

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards accurate and fair prediction of college success: Evaluating different sources of student data. In *Paper presented at the proceedings of the 13th international conference on educational data mining (EDM 2020)*.

Zaffar, M., Hashmani, M. A., Savita, K., Rizvi, S. S. H., & Rehman, M. (2020). Role of FCBF feature selection in educational data mining. *Mehran University Research Journal of Engineering and Technology, 39*(4), 772–778.

Zhang, Y., & Wu, B. (2019). Research and application of grade prediction model based on decision tree algorithm. In *Paper presented at the proceedings of the ACM Turing Celebration conference-China*.

Zhao, L., Chen, K., Song, J., Zhu, X., Sun, J., Caulfield, B., & Mac Namee, B. (2020a). Academic performance prediction based on multisource, multifeature behavioral data. *IEEE Access, 9*, 5453–5465.

Zhao, Y., Ren, W., & Li, Z. (2020b). Prediction of english scores of college students based on multi-source data fusion and social behavior analysis prediction of english scores of college students based on multi-source data fusion and social behavior analysis.

Zohair, L. M. A. (2019). Prediction of student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education, 16*(1), 1–18.

Zong, J., Cui, C., Ma, Y., Yao, L., Chen, M., & Yin, Y. (2020). Behavior-driven student performance prediction with tri-branch convolutional neural network. In *Paper presented at the proceedings of the 29th ACM international conference on Information & Knowledge Management*.

Zulfiker, M. S., Kabir, N., Biswas, A., Chakraborty, P., & Rahman, M. M. (2020). Predicting students' performance of the private universities of Bangladesh using machine learning approaches. *International Journal of Advanced Computer Science and Applications, 11*(3), 672–679.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Saba Batool[1] · Junaid Rashid[2] · Muhammad Wasif Nisar[1] · Jungeun Kim[3] · Hyuk-Yoon Kwon[4] · Amir Hussain[5]

✉ Junaid Rashid
junaidrashid062@gmail.com

✉ Jungeun Kim
jekim@kongju.ac.kr

[1] Department of Computer Science, COMSATS University Islamabad, Wah Campus, Islamabad, Pakistan

[2] Department of Computer Science and Engineering, Kongju National University, Cheonan 31080, South Korea

[3] Department of Software, Department of Computer Science and Engineering, Kongju National University, Cheonan 31080, South Korea

[4] Department of Industrial Engineering, Seoul National University of Science and Technology, Seoul, South Korea

[5] Data Science and Cyber Analytics Research Group, Edinburgh Napier University, Edinburgh EH11 4DY, UK